Year : 2022

The Evaluation of Mobile Device Evidence under Person-Level, Location-Focused Propositions

Spichiger Hannes

The Evaluation of Mobile Device Evidence under Person-Level, Location-Focused Propositions

Doctorate Thesis - Thèse de Doctorat
presented at the
Ecole des Sciences Criminelles
Université de Lausanne
by

Hannes Spichiger
MSc in Forensic Science

Jury:
Prof. Eoghan Casey, supervisor (University of Lausanne)
Prof. Gillian Tully, external expert (Kings College London)
Prof. Charles Berger, external expert (Leiden University)
Prof. Christoph Champod, internal expert (University of Lausanne)
Prof. Rebekah Overdorf, internal expert (University of Lausanne)

President of jury: Prof. Stefano Caneppele

UNIL | Université de Lausanne

Lausanne 2022

The Evaluation of Mobile Device Evidence under Person-Level, Location-Focused Propositions

Doctorate Thesis - Thèse de Doctorat
presented at the
Ecole des Sciences Criminelles
Université de Lausanne
by

Hannes Spichiger
MSc in Forensic Science

Jury:
Prof. Eoghan Casey, supervisor (University of Lausanne)
Prof. Gillian Tully, external expert (Kings College London)
Prof. Charles Berger, external expert (Leiden University)
Prof. Christoph Champod, internal expert (University of Lausanne)
Prof. Rebekah Overdorf, internal expert (University of Lausanne)

President of jury: Prof. Stefano Caneppele

UNIL | Université de Lausanne

Lausanne 2022

UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Monsieur Hannes Spichiger, candidat au doctorat en science forensique, intitulée

« The Evaluation of Mobile Device Evidence under Person-Level, Location-Focused Propositions»

Prof. Stefano Caneppele
Président du Jury

Lausanne, le 28 novembre 2022

Abstract

Location-related evidence recovered from mobile devices is frequently used both in the investigation stages and the trial stage of criminal case proceedings. Due to societal factors and incomplete understanding of the factors involved, these traces are often taken as face value and without a proper separation between the device and the person carrying it. This work proposes a structured Bayesian approach for the evaluation of these traces consistent with existing approaches for physical evidence. The approach not only allows for the evaluation of location-related evidence, but also proposes ways to bridge the Person-Device Gap. In four simulated scenarios, it is shown that the presented approach can be applied on real data and that a likelihood ratio (LR) for digital traces can be obtained.

Des preuves liées à l'emplacement, récupérées d'appareils mobiles, sont fréquemment utilisées dans des procédures pénales. Des causes sociétales et une mécompréhension des facteurs entrant en jeu conduisent à ce que les traces soient souvent acceptées telles quelles, sans considération de la séparation entre l'appareil et la personne le portant. Ce travail propose un approche Bayesienne structuré pour l'évaluation de ces traces. Cette approche est en ligne avec les approches existantes pour des traces physiques et permet d'appröcier les incertitudes liées à la localisation et l'écart personne-appareil. En 4 scénarios, il est demontré que l'approche peut être appliquée à des données réelles et que des rapports de vraisemblances (LR) peuvent être obtenus pour des traces numériques.

Scripts used in this work are available at `https://github.com/HSpichig/Thesis`.

Acknowledgements

One does not write a thesis on its own. Well, one does, but really, there are many people that help, and be it just by listening to a problem that currently seems impossible to resolve. To all those I extend my deepest thanks. This project was possible because of you. In no particular order, and in no way extensive, I especially thank

My supervisor, Prof. Eoghan Casey, for his patience, wise words, encouragement and support through all this. His ability to guide me towards the answer of a question neither of us knew I needed to ask greatly helped me advance.

The members of my thesis committee, Prof. Stefano Caneppele, Prof. Gillian Tully, Prof. Charles Berger, Prof. Rebekah Overdorf and Prof. Christophe Champod for their constructive comments, interesting questions and their courage to be experts in a completely uncharted domain.

Andreas Spichiger, Gaëtan Michelet, Jorina Marti and Maëlig Jacquet for their invaluable comments on earlier drafts of this document.

Elénore Ryser, Timothy Bollé and Jasmin Wyss for enlightening discussions, sometimes at ungodly times of the day.

Lionel Brocard and Thomas Souvignet for help with practical matters and their shared interest in the quirks of location-related traces.

My two anonymous data providers for Scenario 2 for suffering through having to use an iPhone for three weeks.

My sister and my parents that had the patience to listen me talking about location-related traces for the last five years, and took the time to try to understand what I was working on.

My coworkers, friends and family, several mentioned here for the second time, for keeping me sane during the whole process of writing this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Even in the very connected world we live in, many crimes that are committed require the perpetrator to be present on the crime scene. This fact is the basis for many investigative and forensic approaches and manifests itself in two ways: either by gathering traces of presence on the crime scene, or by putting a person at another place during the time of interest. In both of these situations, the aim is to establish locality. The possible existence of such traces is postulated by the Locard's Principle of Exchange (Locard, 1920). Due to the interactions between a perpetrator and the surrounding environment, he or she may leave traces on the scene as well as take elements away from the scene, the intensity of the interaction being key to the quality and quantity of traces that are exchanged. This interaction with its environment is particularly intense for mobile devices, amplifying the quantity of traces that may be recovered. A modern mobile device interacts heavily with its surroundings: sensors measure its orientation, acceleration, time and location, antennas exchange information with nearby devices and cell towers; and integrated cameras have replaced handheld cameras for most amateur, and even some professional, applications. As such, a modern mobile device knows most moments where it is and, due to very detailed logging, it is also often possible to reconstruct its movement at a later stage. It is, therefore, no surprise that data stored on and transmitted by mobile devices has received growing attention from investigators and attorneys alike. (Casey and Turnbull, 2011)

In addition, mobile devices have culturally very much developed as personalized devices. Although it is technically possible to have more than one user profile on Android, this feature is disabled by default since version 5.0 (AOSP, 2020). On Apple iOS devices, only a single user profile is permitted. As such, mobile devices distinguish themselves from other devices that are designed to support multiple user profiles such as personal computers

and servers. A multitude of technical applications work on the assumption of "one user one device," such as two factor authentication for personal accounts (Rogers, 2011), targeted advertising (Knowlson, 2003), and device based traffic detection (Thiagarajan et al., 2009). These applications functioning in general without larger issues indicates that the assumption is sufficiently supported for large scale applications to work acceptably well. Even though singular counterexamples exist, such as a German artist provoking artificial "traffic" on Google crowd-sourcing services with a trolley full of mobile devices (Shammas, 2020), the assumption is supposed to be sufficiently strong that multiple countries based the core concept of COVID-19 tracking applications on a mobile device being associated with a particular person (Troncoso et al., 2020; FOPH, 2020).

No formal studies exist whether this assumption remains valid in criminal investigations. However, it is common practice to call upon practitioners to exercise caution (FSR, 2020), indicating that a significant part of the forensic science community considers the distinction between device and person to be essential. It is unclear to what degree the distinction between device and person is actively considered in court. The UK Forensic Science Regulator's guideline to cell tower evidence specifically addresses issues of device location or general ownership, without directly discussing the possession at a given moment (FSR, 2020). Analysis of cases involving mobile device location evidence revealed a lack of discussion of this person-device distinction, leaving the assumption that it was not contested (Kuhn, 2018), sometimes enforced by the circumstances of the case (Circuit Court of Albermarle County, 2015). In cases where such evidence was rejected or deemed insufficient, it was done so on other grounds (Swiss Federal Court, 2019; Poser, 2017). Cases are known where the ownership of a mobile device was contested by the accused, and respective evidence was presented, including which cell towers the phones connected to. Both of these cases were, however, on the level of general ownership and the court's decision focused on the quality of the evidence given, not the reasoning itself (EWCA, 2017, 2020). A few cases are known to the author where it was questioned who was the author of a message written on a device or using a social media account. In at least one case, the case was dismissed based on the prosecution not having met the burden of proof to show general ownership in the account (Superior Court of Pennsylvania, 2018). To identify the author of a specific message, linguistic analysis such as stylometry have been used. For example, two independent analyses claim to have identified the two authors between Q-Anon messages as Ron Watkins and Paul Furber (Kirkpatrick, 2022).

In this thesis, it is postulated that, due to the aforementioned high intensity sensory properties of mobile devices, it is not only possible to find traces

supporting the location of the device but also traces that give indications about the physical identity of the person holding the device at the moment of interest. Therefore, logically, it should be possible to make inferences about the location of the physical person at the moment of interest.

To this day, different approaches are applied in crime laboratories and police forces around the world to address the issue of a person's location based on smartphone traces. The following approaches were mentioned to the author in personal exchanges between 2019 and 2022 and represent the practice at the moment the exchange took place. It is well possible that these practices were in the mean time adapted, so no indications are given as to which services are using these approaches.

- Several services consider their job done once the data is recovered from the device and a report containing all the data is generated and passed on to investigators or the court. This approach is highly problematic, as the investigators and prosecutors treating the case are unlikely to have the specialised knowledge required to evaluate or question the recovered traces.

- In multiple services, conclusions are systematically presented as the location of the device, leaving it to the prosecutor or investigator to link the device to the person.

- Some services systematically write conclusions as the location of the person, assuming the person using the device at time t is the general user of said device. This assumption is made explicit and it is indicated that information contrary to the assumption may change the conclusion.

- Some services do investigate the ownership of the device, however only on a general level and not at a specific moment in time. This approach is, for example, indicated in the forensic science regulator for England and Wales' guideline to the presentation of cell tower evidence (FSR, 2020).

With growing literacy of lawyers and judges in the domain of digital evidence, it should be expected that this «person-device gap» will rightfully be raised more frequently.

One core aspect of the issue is handling uncertainties linked to digital traces through evaluation. This is currently not a practice that is widespread. Of all services talked to, only the NFI indicated conducting evaluation of the

results on a regular basis, based on a Bayesian approach still under development (cf. Bosma (2022)). The Vaud cantonal police (CH) have conducted such an evaluation once so far (Bassi and Scoundrianos, 2022). Both these approaches are based on likelihood ratios (LR).

LR approaches have been criticised for use on digital traces, with one author going as far as to state that «*achieving a scientific mechanism for quantifying [uncertainty in] digital evidence may not actually be feasible due to the nature of digital evidence*» (Horsman (2020), p.1). Whilst not too often stated in published literature, this manner of thinking about digital forensic science is widespread among practitioners, considering digital traces to be fundamentally different from classical traces. In an answer to the article by Horsman, this position of «digital evidence exceptionalism» (Biedermann and Kotsoglou, 2020) was strongly questioned, going as far as to question the use of digital traces if Horsman's assumption were to be true: «*If this is what (digital) forensic science is or aims at, then it is difficult to see how it can meaningfully serve the needs of factfinders in the pursuit of justice*» (Biedermann and Kotsoglou (2020), p.272). In this work, it is shown that Horsman's assumption is in fact erroneous, by presenting a means to quantify the uncertainty of selected pieces of digital evidence. By doing that, it is also shown that digital forensic science has the potential to be of use to decision makers and that it should not be considered as exceptional in regards to evaluation.

Additionally, this dissertation provides the following novel contributions:

- The problem is analysed in detail and structured.
- Means to close the Person-Device gap are discussed.
- A Baysian approach to address uncertainties resulting from the Person-Device gap is presented.
- A Bayesian network for the evaluation of location-related traces on person-level is proposed, studied and tested.
- The possibility to distinguish between two users on the same device is demonstrated.
- It is shown that it is possible to quantify uncertainties linked to digital traces.
- A rudimentary approach for the evaluation of cell tower evidence is proposed.
- An approach for the evaluation of device localisations is presented.

- Issues related to the application of this approach in real world cases are discussed.

In this work, the Person-Device Gap is frequently discussed. To simplify the discussion of this issue, a taxonomy of propositions is introduced, allowing to more easily state on what level an opinion is expressed.

Definition 1. *The* **subject-level** *of a proposition describes the entity at the center of the action described by the proposition.*

The subject level is a hierarchical taxonomy. The levels are stated in Table 1.1. It is based on the identity-concept presented in (Jacquet-Chiffelle, 2008) and allows to distinctively describe levels that may be encountered when working with object evidence of any type.

Level	Explanation	Example
Person	The subject is a physical identity.	P: Person X is responsible for the hack.
User	The subject is a tautological virtual identity. («The user of the object / device»)	P: The user of machine Y is responsible for the hack.
Account	The subject is a virtual identity.	P: User account «Z» was used in the hack.
Object / Device	The subject is an object / device.	P: Machine Y was at the origin of the hack.

Table 1.1: Different subject-levels and their meaning

This thesis is structured as follows: The problem is defined and structured in the remainder of Chapter 1. Existing literature is discussed in Chapter 2. Chapter 3 develops and explains the framework used in this work and studies its behaviour. Chapters 4 through 7 each contain a scenario illustrating the use of the presented framework and showing its applicability to real world scenarios:

- A scenario with location-focus on device-level in Chapter 4.
- A scenario focused on the identity of the user in Chapter 5.
- A scenario with location-focus on person-level with direct evidence of usage in Chapter 6.

- A scenario with location-focus on person-level with indirect evidence of usage in Chapter 7.

Finally, a conclusion is reached in Chapter 8

1.1 Motivation and Problem Statement

An object is considered a device, once «virtual components are added to allow the object to act on both the physical and virtual world» (Casey et al., 2020b). That definition of a device is used as a basis to define a mobile device:

Definition 2. *A* **mobile device** *is a transportable object that can act and sense both on the physical and virtual world.*

As such, mobile devices comprise not only smartphones and tablets, but also mobile IoT objects of varying forms, such as drones, smartcars or wearables. Due to their heavy interactivity with their environment, such devices are known to generate large quantities of data (Casey et al., 2020b). Mobile devices, in turn, are of particular interest, because they generally also create large quantities of data about their position. The extension to the location of the user of the device is often insinuated, potentially leading to the common mistake of assuming that the location of the device is always the same as its owner. In fact, there are multiple circumstances in which this assertion is wrong: the device was elsewhere, a different person was carrying the device, or the device belongs to a different person altogether. The interactivity of mobile devices with their environment is, however, to such a high degree that it is clearly conceivable that evidence supporting the link between the device and the person, and consequently between the person and the device, can be found on the device. This work aims to establish whether this assumption is true.

Research Hypothesis 1. *It is possible to gain, from a mobile device, for a given moment in time, relevant traces about where that device was and who was using it, allowing an expert to express an opinion on a person's whereabouts, supported by a structured reasoning process.*

Figure 1.1 shows a visualisation of the relations articulated in Research Hypothesis 1.

These traces are subject to numerous inherent uncertainties and imprecisions. Even if issues related to recovery and interpretation of the data are

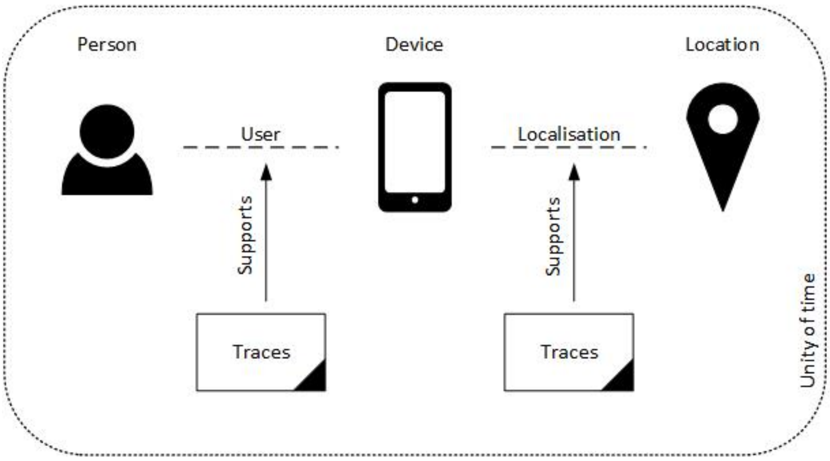


Figure 1.1: Visualisation of Relations Articulated in Research Hypothesis 1

ignored[1], there is uncertainty related to the position as well as resulting from the separation of the device and the user. Current approaches attempting to resolve this issue rely heavily on what an expert considers to be justifiable and defensible, without being supported by data. This is not a problem, per se, but there is widespread consensus that structured approaches supported by measurable data would be an improvement (Biedermann and Kotsoglou, 2020). Biases in forensic science in general, but with digital traces in particular, can have a high degree of diversity, carrying an increased risk of misjudgements that may not even be based on bad scientific foundations, but simply because the assessment of the strength of evidence was clouded (Sunde and Dror, 2019). These biases can be mitigated by explicit and structured assessment of evidence based on measurements. Such approaches are expected to be supported by the ISO/CD 21043-4 standard on Forensic Science (Interpretation) currently under development (ISO/TC 272 Forensic sciences) and may be required by judges or regional standards (e.g. (FSR, 2021) for England and Wales). Indeed, multiple cases exist in which the reliability of location evidence was considered insufficient and thrown out by judges (Gavin, 2017; Swiss Federal Court, 2019), where adapted evaluation could have led to the evidence being helpful for the case. There does not currently exist a way to structure and combine uncertainties into an overarching result in cases where mobile device evidence is used to pinpoint the location of a person. This work aims to address this issue and provide the groundwork to resolve it.

[1]These sources of error should not be ignored, however, this is out of the scope of this work.

Research Hypothesis 2. *Traces from a mobile device can be evaluated in a logically consistent manner under a pair of location-focused propositions with a physical person as a subject.*

The overall aim in this work is to evaluate traces under propositions of the following form:

P_1:	Person A was at location X at time t.
P_2:	Person A was at location Y at time t.

There are subtleties linked to these propositions. A detailed understanding of them can help with the understanding of the issues at hand. First, the propositions only differ in where person A was at time t. Neither the identity of person A, nor the moment in time is doubted. Second, person A is a physical person. As stated, the observed location-related trace is only directly connected to the device. A reconstruction linking the physical person to the location is proposed in this work, allowing to conduct evaluation on the person-level. Third, the hierarchy following Cook et al. (1998) should be inquired, as it may be counter-intuitive to some. With classical object traces, evaluation on a person-level was inherently linked to at least activity-level propositions, as object traces by themselves did not allow to reach such a conclusion Cook et al. (1998). As already argued, this is no longer the case. As neither relevance, a central element of offence-level evaluation, nor transfer and persistence, central elements of activity-level evaluation, are investigated here, it is concluded that these propositions are on source-level, despite talking about the location of the person and not the device. One way this can be understood is by looking at the element of interest: the location. As it is aimed to distinguish between two locations, traces inspected are generated by the environment at those locations, assumed to be distinguishable. In other words, the question is whether location X or location Y is the source of the observed trace, and not the person or the device.

Based on this understanding of the propositions[2], a structured and formal approach for evaluation can be established.

1.2 Structuring the Problem

As already stated, the problem can be split up in two steps. As visualised in Figure 1.1, 1) the device needs to be localised, and 2) the usage of the device needs to be established at the moment of interest t. This two-step approach

[2]The present analysis is valid for the overarching pair of propositions. In this work, the propositions are treated in sub-hypotheses, which may have different classifications.

is reproduced at a later stage in this work on the level of the likelihood ratio (LR). The scenario has basically two extreme states. If it is categorically known that the device was used by person A at time t, the location of the device is identical to the one of the person and related uncertainties are the same as well. In this situation, the LR on the person-level is identical to the LR on the device-level. The second extreme, if it is known that the device was not used by person A, the location of the device cannot give any relevant information about the whereabouts of person A and the LR becomes 1. For all other situations, the person-level LR will be found somewhere on the spectrum between those two extremes, based on the impact of the evidence of usage. In other words, the location-related evidence gives a range of possible person-level LR and the evidence on usage indicates where in this range, the person-level LR lies exactly. This process is visualised for LogLR in Figure 1.2[3].

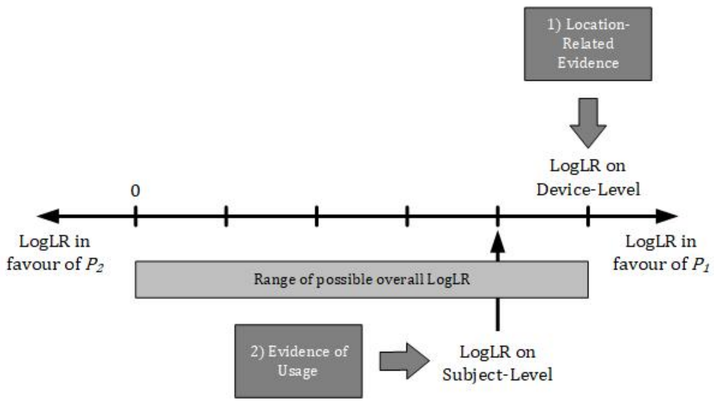


Figure 1.2: Process of moving from a device-level level LogLR to a person-level LogLR.
First, the location-related evidence gives a range between 0 (signifying irrelevant evidence, corresponding to an LR of 1) and the device-level LogLR, second, the evidence of usage indicates where in this range the overall value is situated.

In more detail, the problem is structured as a reasoning tree, visualized in Figure 1.3. Each node of the tree represents its own forensic question, complete with at least a pair of competing propositions and a non-extensive list of traces that have the potential to be relevant for the answering of this particular node. At the end of each node stands an evaluative conclusion

[3]The process is equivalent for LR. The visualisation with LogLR is chosen for the axis to be symmetrical.

for this particular question. Arrows feeding into a node represent evaluative conclusions reached in earlier nodes.

This reasoning model is intentionally independent of the chosen evaluative method. Even though this work applies a Bayesian approach using Likelihood Ratios, the reasoning structure can be used just as well for any other method cited in Section 2.1. As such, it is still valid in cases where the answer to the question of one node is categorically known. In this case, all nodes leading up to this node can be ignored and only the rest of the tree has to be treated.

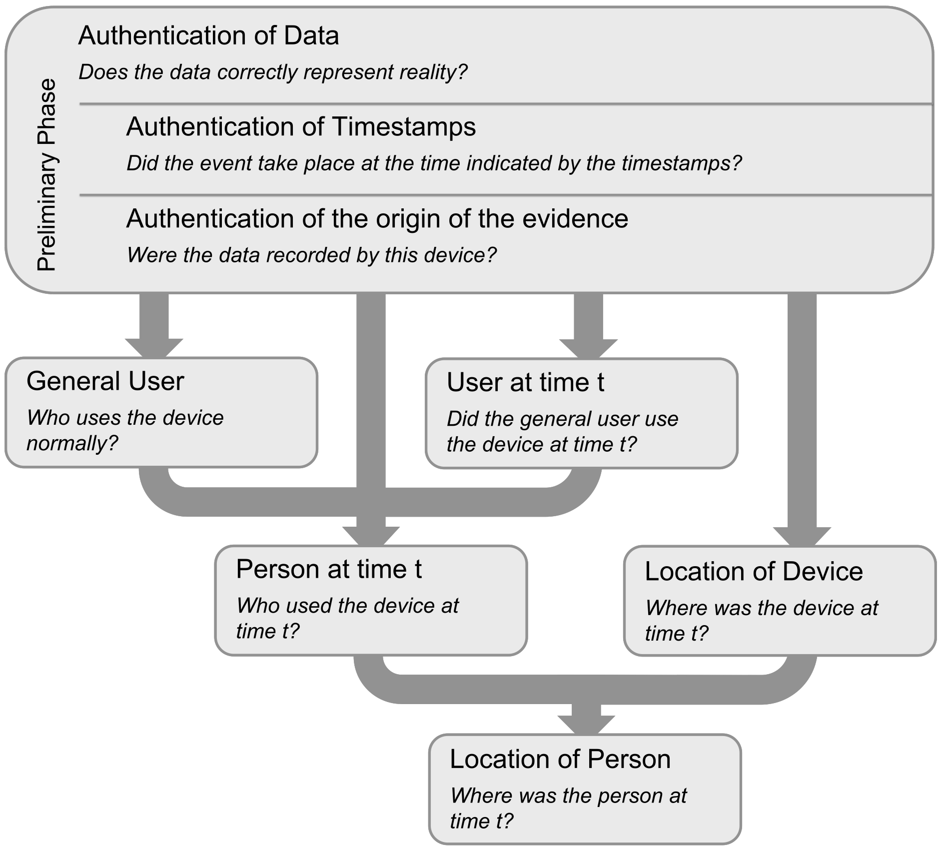


Figure 1.3: Underlying reasoning structure of the problem.

Preliminary Phase

Can the traces be used for further analysis?

The preliminary stage consists of a technical verification of the traces. Conclusions on digital evidence may only be drawn when the following three aspects have been verified:

Authentication of the Trace: The data of interest has been correctly recovered from the device and is correctly interpreted by used tools. This aspect also takes into account issues that may result from the state of the device at the moment of the creation of the traces such as geo-spoofing or technical errors. The verification of trace authenticity remains a challenging issue that is likely to be the topic of many research projects in the following years. Current approaches favour the correct application of methods and internal consistency of the traces to affirm a low risk of authentication not being met. (Kuntze et al., 2012; Arshad et al., 2018)

Authentication of the Timestamp: The entirety of the reasoning within this work is for a specific moment in time. Consequently, only traces in temporal proximity to this moment can have relevance for the answering of the question. Therefore, only traces where the correctness of the timestamps can either be reasonably assumed or verified can be taken into account for further reasoning. To the knowledge of the author, the authenticity of the timestamp can currently only be assessed by a lack of contrary evidence, comparison with other sources containing traces of the same information as well as the verification of internal consistency.

Authentication of the Origin of the Traces: As the argument for using mobile device traces is based on the strong interaction of the device with its environment, evidence can only be relevant if it is generated by the device itself. Due to the inter-connectivity of modern mobile devices, it is not sufficient to have extracted data from a device to ensure it was also created by it. It is for example known that Apple services send the location of WiFi access points and cell towers to the device independent of whether the device was ever connected to them or has even detected their presence (Forensic Focus, 2011). This data can be found in device extractions and may be mistaken as location-related evidence. The question of origin is particularly pronounced for media files, as they are frequently exchanged by users through messaging apps. Logically, they can only give information about a particular device, if they have been recorded by it and not transferred to it from another device. On mobile devices, the location of storage, the file name and the metadata are generally indicative of this. For images, more in-depth analysis possibilities such as PRNU exist (Houten et al., 2011; Spichiger, 2017).

These verification steps are required for all digital traces, independent of the context of their use. Since this stage is of quite advanced complexity, it would greatly surpass the scope of this work. It is therefore assumed that the

authenticity and the source can be affirmed for all available traces. Section 3.1 will briefly discuss how non-categorical statements from the preliminary state would have to be introduced in the Bayesian network, but none of the cases will work with such situations and only treat traces assumed to be categorically reliable.[4]

Identification of the General User

Who is the person generally using the device?

In this stage, the physical person using the device normally is identified. As mentioned in the introduction, it is assumed by society that this is one person, even though there is no particular reason that this would always be the case. Identifying who generally uses a device is common practice for many Digital Forensic practitioners. Traces generally considered in those cases contain but are not exclusive to usernames, selfies, interests and activities, as well as contacts.

Identification of the User at Time t

At time t, was the person using the device the general user?

This stage is somewhat abstract: The aim is to authenticate whether or not at time *t*, the same person is using the device as the person that normally uses the device. This process is to be considered an identification process in which the link between the «user at time *t*» and «the general user of the device», is established. Indeed, in absence of the previous stage, the conclusion of this stage is only whether the same person as normally is using the device, without expressing anything about the physical identity of this person. Traces giving information about this stage are considered to be found within the activities on the device. It is postulated that a change in the person using the device would result in a change in the activity traces recorded by the device. In some cases, the conclusion of this stage may be that no one was using this device at time *t*. In this case, the localisation of the phone becomes irrelevant for the localisation of anyone.

Identification of the Person at Time t

Who is the person using the device at time t?

[4]This is not be interpreted as a statement, that these traces *should* be considered this way. It is the sincere conviction of the author that none of the above should ever be assumed without having concrete evidence. Consequently, resolving the issue of traces authentication will likely be one of the most pressing issues of the years to come.

This stage combines the results of the previous stages (If the general user is identified and the general user is the one using the device at time t, then the person using the phone at time t has to be the general user.) but its conclusion can also by itself be directly supported by traces. There may for example be media recordings containing biometric information about the person using the device at the time of the recording.

Localisation of the Device at Time t

Where is the analysed device at time t?

This stage comprises the localisation of the mobile device, as it is classically done by digital forensic experts. Existing literature relevant for this stage of reasoning is cited in Section 2.2. At the end of the stage stands an evaluative statement on the location of the device at time t, independent of the person in which possession it was at this moment.

Localisation of the Person

Where is the person of interest at time t?

This final stage of the reasoning process is quite simple, as it is just a logical combination of all previously reached conclusions. If the person of interest was in possession of the device at the moment of interest, the position of the person is the same as the one of the device. Otherwise, if the person was not in possession of the device, the position of the phone is irrelevant to determine the position of the device.

1.3 Practitioner Considerations

The approach presented in this work aims to be applicable to real world situations. However, practitioners should consider whether expressing an opinion on the location of the person of interest is actually adapted for their case. The following situations may indicate that the presented approach is not well adapted:

The person is known to have had the device: If, for whatever reason, the person of interest is known to have been in possession of the device during the time in question, or this is not contested, the location of the person has to be the location of the device, including associated uncertainties. In such situations, the LR on person-level becomes equal to the LR on device-level (as detailed in (Casey et al., 2020a)), and the entire reasoning about the user of the device can be left out.

The person is known to not have had the device: If the person is known to not have been in possession of the device during the time in question, information about the device's location cannot be relevant for the question of the person's location. Time spent on analyses in that regard is better spent otherwise.

High uncertainty about usage: If the observed evidence on the user is very weak, it may be better to express an LR on the device-level, whilst clearly stating the limitations of the conclusions. That way, the analysis can still be of use to the court if they manage to place the device in the hand of the person of interest. Such a situation was encountered in the scenario presented in Chapter 7.

The location of the device is clear: If the location of the device is clear, or the LR on the device-level approaches infinity, the overall LR is likely to become quite volatile, changing orders of magnitude based on small changes in the likelihoods of usage. In such a situation, it is recommended to address the two issues separately and provide LR for both questions.

Situations with large LR on device-level may result from large distances between the observed location and one of the propositions. This is more likely to occur if both propositions are far apart. In the past, practitioners have based their reasoning on the distance between the two locations indicated in the propositions. This information is not relevant for evaluation and should not be taken into account; it is the distance of the device from each location in each proposition that is relevant to the evaluation as demonstrated in this work.

Chapter 2

Existing Work

The three core aspects of this work are evaluation, evidence in relation to location, and evidence in relation to user identity. Existing work for each of these aspects is laid out in this chapter. Section 2.1 studies literature on evaluation as a methodology to quantify uncertainty in forensic evidence. As there is very little discussion of the evaluation of digital evidence, with Subsection 2.1.1 focusing on this topic aims to be as exhaustive as possible. Location-related evidence and existing literature are discussed in Section 2.2. An overview of literature addressing user identification is given in Section 2.3.

2.1 Evaluation of Forensic Evidence

For many, the most visible aspects of forensic science are the technical examinations, chemical treatments or, for the layperson almost magical, applications revealing so far invisible traces. But just as important is a much later stage, where the found and treated traces are put into context with the case and the competing propositions in the latter. This phase is called Evaluation. In classical comparative domains of forensic science, the evaluation is the third stage in the so called ACE-V process, where the meaning of observed correspondences and differences is assessed (Huber, 1959). A more general definition of Evaluation is given by the OSAC committee for digital and multimedia evidence:

Definition 3. Evaluation: *Produce a value that can be fed into a decision process. (Pollitt et al., 2018)*

The term «value» from this definition is understood in a broad sense and does not have to be numerical. Any at least ordinal classification system can be used as an output of an evaluation process. In the context of forensic science, evaluation is often associated with the attribution of a likelihood ratio,

in short LR. First proposed in (Finkelstein and Fairley, 1970) for identification evidence, the concept was published following a series of court cases in which probabilities were presented incorrectly. Finkelstein and Fairley base their approach on the theorem for conditional probabilities, attributed to Reverend Thomas Bayes[1] (Bayes and Price, 1763). Bayes' theorem (Formula 2.1) describes how the probability of an event can be calculated if another event is known to have taken place.

$$Pr(P|E) = \frac{Pr(E|P)Pr(P)}{Pr(E)} \tag{2.1}$$

Applied to court proceedings, the theorem formalises how the observation of a piece of evidence (E) influences the belief held into the truth of a proposition (P). As generally at least two concurring propositions are disputed in front of a court, the theorem is often represented in its *odds form* (Formula 2.2), the division of two conditional probabilities.

$$\frac{Pr(P_1|E;I)}{Pr(P_2|E;I)} = \frac{Pr(E|P_1;I)}{Pr(E|P_2;I)} \times \frac{Pr(P_1|I)}{Pr(P_2|I)} \tag{2.2}$$

In its odds form, the theorem can be read as a mathematical representation of the judicial process: On the right, with the *prior odds*, the probabilities of the presented propositions before presented evidence is taken into account ($Pr(P_n|I)$), purely based on the relevant information about the case (I). Next, the power of the evidence to sway the odds in one direction or the other is represented through the *likelihood ratio* (LR). It is the quotient of the probabilities to observe the traces alternately *assuming* the propositions to be true ($Pr(E|P_n;I)$). In other words, a universe is imagined, in which the proposition is known to be true as a fact. The probability to observe the presented findings in this universe is then assigned. Finally, by combining the two previous terms, the *posterior odds* on the far left are obtained. They represent the belief in the propositions after having integrated the information about the observed evidence and are the quotient of the probabilities that the respective propositions are true given the observed evidence ($Pr(P_n|E;I)$). If multiple pieces of evidence are presented, the process can be applied in a cyclical manner where the posteriors of the previous step become the priors for the next one, if the propositions remain the same and the later pieces of evidence are evaluated conditioned by the evidence evaluated earlier. In a

[1]The essay was published posthumously by Richard Price who found it amongst Rev. Bayes' papers.

court setting, it is generally considered to be the forensic scientists role to only talk about the likelihood ratio. The narrow nature of this role is due to insufficient knowledge about the priors as well as it quite simply not being the task of the expert to talk about them and often require knowledge outside of the experts area of expertise (Thompson et al., 2013). The very same framework can however also be used to support decisions in an investigative or pre-analysis phase. In this case, the forensic scientist may indeed be well placed to address the entirety of the process, including posterior odds, if the necessary information is available[2] (Ryser et al., 2020; Baechler et al., 2020). To put numerical LR in context, experts have often chosen to present their results using a verbal scale, generally based on a logarithmic grouping of values in verbal categories (ENFSI, 2010; Marquis et al., 2016). This concept of logarithmic quantification can be found in an approach as well, whereas the logarithm of the LR is presented instead of the LR itself. The multiplicative property of LR then becomes an additive one, which has been argued is more intuitive to people without a strong background in probability theory (Pierce, 1877; Aitken et al., 2018). Addressing problems of increasing difficulty, it has become commonplace to use Bayesian networks, or Bayes Nets for short, to calculate LR. A Bayes Net is a directional, acyclic graph where variables are represented by nodes and dependencies between the variables are represented by edges (Taroni et al., 2014). Each node has assigned a table of conditional probabilities that can be updated dynamically if states for some variables become known. Bayes Nets are particularly of interest when propositions on activity- or crime-level are considered or multiple pieces of interdependent evidence are observed. Especially with complex problems, they have the advantage that, in addition to the calculations they do, Bayes Nets follow in their structure generally the structure of logical reasoning and can therefore also be used to talk about the influence of a variable on the LR as a whole (Taroni et al., 2014).

2.1.1 Evaluation of Digital Evidence

For a long time, and in some circles still today, digital evidence has been considered as purely factual. For example, an expert may state that «the device was at place X at time t» without taking into account uncertainties linked to the found data. Other used terminology «the data is consistent with», whilst technically correct, fails to acknowledge alternate possibilities and the strength of the evidence in light of them (Bunch, 2014). With dig-

[2]This is more likely to be the case early in the investigative process, or if the question addresses purely technical aspects.

ital evidence becoming more integrated with Forensic Science as a domain with principles valid for all subdomains, reflections about the presentation of evidence are starting to catch up with classical Forensic Science domains.

The first known instance of an author explicitly addressing the inherent uncertainty of digital evidence and attempting to quantify that uncertainty can be found in (Casey, 2002). The proposed approach, called the C-scale, is an ordinal scale where each level has a clear description of what requirements have to be met for a level to be attained with a strong focus on protection against tampering. This initial C-scale (shown in Table 2.1) had multiple issues, notably that the description of the levels sometimes stopped additional evidence from having an impact on the classification level. Also, the scale was a hybrid between evidence-focused and proposition-focused, which meant that it was very difficult to assess whether its use was appropriate or not. The C-scale was reworked in (Casey, 2020) to address these exact issues. In its new, adapted form, the C-scale (shown in Table 2.2) is purely focused on evidence and allows for intermediary steps, as such allowing a more differentiated evaluation of the evidence. The C-scale can be categorized as a proto-Bayesian approach that can be used following a Bayesian logic, without having to ascribe probabilities (Ryser et al., 2020). As such, the C-scale, especially with the indicators associated to the different values, has the potential to serve as an intermediary for practitioners wanting to follow a Bayesian logic but not yet feeling comfortable with the use of numerical probabilities.

Some attempts were made to resolve issues with digital evidence using Bayesian Networks. Most of them suffer from fundamental shortcomings relating to the use and understanding of the Bayesian Approach and framework. Kwan et al. (2008) present a Bayes Net to address the question of whether a computer seized in a BitTorrent case was the initial seeder for a pirated file. This work used the Bayes Net to provide posterior probabilities, not to obtain an LR. Prior probabilities are assumed as uniform, but their origin is not discussed[3]. The network also includes «uncertain» states for some nodes, all evidence was considered to be independent from each other without justification of the latter, and conditional probabilities within sub-hypotheses are assigned without justification and in a way that the main hypothesis could be true, even if all sub-hypotheses were found to be false. One interesting aspect of the work is that values are assigned based on a weighted mean from opinions given by experts in the field. A sensitivity analysis of the same Bayes Net is conducted in Overill et al. (2010) varying

[3]This would allow to easily obtain an LR from the network if it were to be used in such a way.

Certainty Level	Description/Indicators	Commensurate Qualification
C0	Evidence contradicts known facts.	Erroneous/Incorrect
C1	Evidence is highly questionable.	Highly Uncertain
C2	Only one source of evidence that is not protected against tampering.	Somewhat Uncertain
C3	The source(s) of evidence are more difficult to tamper with but there is not enough evidence to support a firm conclusion or there are unexplained inconsistencies in the available evidence.	Possible
C4	Evidence is protected against tampering or multiple, independent sources of evidence agree but evidence is not protected against tampering.	Probable
C5	Agreement of evidence from multiple, independent sources that are protected against tampering. However small uncertainties exist (e.g., temporal error, data loss).	Almost Certain
C6	The evidence is tamper proof and unquestionable.	Certain

Table 2.1: Initial C-Scale according to Casey (2002).

the probabilities indicated by domain specialists from the lowest indicated value to the highest and observing the impact of missing evidence and shown to be relatively robust against these changes (Overill et al., 2010). However, as the Bayes Net is reused without addressing the flaws in the original design, there is limited use in their results. In particular, the «uncertain» states are susceptible to mitigate these changes quite a bit. The same network is used as an illustrative example in Tse et al. (2012) where a methodology to construct Bayes Nets is presented (Tse et al., 2012).

In 2012, Overill and Silomon proposed an approach based on a complexity estimation to approximate the likelihood of cybercrimes being committed using a Trojan horse program. The underlying argument, that more complex operations are less likely to be performed inadvertently, seems pertinent to some extent. However, the authors fail to provide a rationale for the persistence of the inverse proportionality between the proposed complexity measure and the probability of the proposition given the evidence. As with

C-Value	Verbal level	Illustrative indicators
C0	Erroneous/Incorrect	Evidence contradicts known facts. (Extreme dissonance of observations in light of the hypothesis)
C1	Extremely weak evidence	Evidence is highly questionable (very strong dissonance of observations in light of the hypothesis).
C2	Very weak evidence	Only one source of evidence that is not difficult to tamper with.
C3	Weak evidence	The source(s) of evidence are more difficult to tamper with but there is not enough evidence to support a firm conclusion or there are unexplained inconsistencies (dissonance) in the observed evidence in light of the hypothesis.
C4	Strong evidence	The source(s) of evidence are much more difficult to tamper with evidence from multiple, independent sources (strong harmonious observations in light of the hypothesis).
C5	Very strong evidence	The source(s) of evidence are very much more difficult to tamper with and evidence from multiple, independent sources (very strong harmonious observations in light of the hypothesis). However, small uncertainties exist (e.g. temporal error, data loss).
C6	Extremely strong evidence	The evidence is tamper proof (or tamper evident) and extremely strong harmonious evidence in light of the hypothesis unquestionable

Table 2.2: Adapted C-Scale as presented in (Casey et al., 2020a)

previous publications by the same authors, the paper positions itself on expressing posterior odds, without discussing the implications of this. The authors finally propose what they presume to be an "upper bound plausibility" for the Trojan horse defence strategy (Overill and Silomon, 2012). In Overill et al. (2013) the probability of inadvertently downloading a given number of child sexual assault materials (CSAM) amongst a larger number of legal pornography is modeled. In this work the authors only assess the probability of the evidence under the defence proposition. In addition, a number of assumptions are made to model the problem as an element selection problem. In reality, this problem is likely to be more complex and no evidence is provided to show that the stated assumptions hold up in reality. An overview of the papers by Overill and Kwan was published in (Overill and Collie, 2021) without discussing the issues with those papers discussed above.

An ostensibly functional approach has been proposed by Biedermann and Vuille in 2016. They analyse the use of evidence by the Swiss Federal Criminal Court in a case of attempted homicide by the use of explosives. In the case, cell tower evidence was discussed and the authors present a Bayes Net to assess the findings. That network shows similarities to parts of the network used later in this work. Casey et al. (2020a) have criticised some of the conclusions reached in (Biedermann and Vuille, 2016). Indeed, in the ruling the courts states that, «at the moment of the act, no mobile device belonging to the appellant could be located at the scene»[4]. This poses a problem as the network Biedermann and Vuille present is not built to take this possibility explicitly into account. Their node E, representing the findings has only two states, the phone connecting on the antenna in question and the phone not connecting to the antenna in question. For the network to provide a sensible answer in absence of any observation, the second state should be split up to take into account the possibility of a connection elsewhere and no connection at all, as these two cases drastically differ in their evaluation. Overall, the LR of absence of localising information is likely to be 1 (Casey et al., 2020a). A more general structured approach for the evaluation of location related evidence was provided in (Casey et al., 2020a). Their work presents the evaluation process as part of a decision process independent of the type of evidence and the chosen way to communicate results. Like this work as well, (Casey et al., 2020a) is limited to the issue of precision and assumes the evidence to be recovered and represented properly. In structuring the

[4]Translated from German. The original sentence is «[...] zum Tatzeitpunkt keine Verkehrsdaten der Mobilfunkgeräte des Beschwerdeführers am Tatort geortet werden konnten.» (Swiss Federal Criminal Court, 2014)

issue, they discuss typical propositions encountered in those cases and what the consequences of those propositions are. A real world case from the Vaud cantonal police in Switzerland, in which this approach was applied was presented by Bassi and Scoundrianos (2022). Based on simulations reproducing what was claimed to have happened by the prosecution and the accused in a homicide, they expressed an expert opinion in the form of LR. The question of whether two mobile devices were used by the same person was addressed by Bosma et al. (2020) and De Bie (2022). Their LR approach is based on comparing two series of cell tower connections and estimating the probability of observing the present similarities with the same or two different persons using the devices respectively (Bosma et al., 2020; De Bie, 2022). The same group of researchers is also working on a model to evaluate connections to a given cell tower under a pair of location related propositions. Their model is based on large quantities of data measured from police patrol cars. Combining data from distance and angle to the cell tower, they present a model that, once sufficiently complete, may have the potential to provide LR-values for cell tower evidence without having to conduct measurements in the field if sufficient knowledge of the terrain and cell tower configuration is available. This ongoing work was presented in (Bosma, 2022).

Galbraith et al. (2020) propose two approaches to evaluate whether two sets of geolocation-data obtained through geofencing warrants were created by the same person. With both likelihood ratios obtained through kernel-density estimation and a similarity-score based approach, they obtain numerical LR to support a decision whether or not the data should be supposed to be from the same individual (Galbraith et al., 2020). Their approach is inherently linked to an investigative phase, as the common-source proposition can then be tested with user data that may be obtained through a warrant. In their tests, the score based LR yields a higher rate of misleading evidence. As there is however no calibration of their system, there is a chance that there may be improved upon that. There is no discussion whether two persons travelling together may cause an unexpected similarity through this type of data.

Skepticism has been expressed towards the use of probabilities for evaluating digital evidence, mostly because for many cases, it seems to be challenging to find data supporting the assignment of any probability (Horsman, 2020). That this is a misconception has notably been pointed out by Biedermann and Kotsoglou (2020). Referencing the ENFSI guideline for evaluative reporting in forensics science (ENFSI, 2010), they indicate the possibility to draw from experience and training in the absence of relevant databases or structured evidence (ENFSI, 2010; Biedermann and Kotsoglou, 2020). With digital and biometrical evidence, the use of automated tools may have to be

considered in addition as well (Bollé et al., 2020a). The usefulness of the underlying logic of a likelihood ratio, even in absence of probabilistic models, was also emphasized in Tart (2020) and Tart et al. (2021) in regards to cell site analysis. A real world example of expert opinion framed in a Bayesian manner was presented in Bollé et al. (2020b).

An alternative approach for evaluation was proposed by Horsman in 2020. His «Digital Evidence Certainty Descriptors» (DECDs, cf. Table 2.3) are similar to the original C-scale (Casey, 2002) in that they provide an ordinal scale of levels depending on how strong the expert believes a proposition to be true Horsman (2020). His descriptors aim to qualify what in a Bayesian approach would correspond to posterior probabilities. They are not conceived to compare multiple scenarios against each other and which could signify that their integration in a logical approach could prove complicated. Additionally, as the lowest and highest level require absolute certainty, something that is close to impossible to reach, the scale comes essentially down to a four level scale where the middle levels are not mutually exclusive. Also, the scale does not provide a level for situation where observations suggest that the scenario is not what happened, but where an absolute level of certainty cannot be reached. As such, the scale does not meet the requirements to address the issue of uncertainty in an adapted manner and their use is not recommended. These aspects are extensively discussed in (Biedermann and Kotsoglou, 2020).

Level	Descriptor	Significance
1	*Conclusive Fact*	The scenario is known to be true with absolute certainty.
2	*Persuasive*	All observations are "consistent" with the scenario, however it cannot be categorically proven.
3	*Conceivable*	No disagreeing observations were made, however, multiple scenarios remain equally possible.
4	*Insufficient Information*	There is not enough information available to reach a conclusion.
5	*Implausible*	The scenario cannot be disproved, but none of the available data suggests it did.
6	*Impossible*	Given the functionality of the observed device, the proposed scenario is impossible.

Table 2.3: Digital Evidence Certainty Descriptors (DECDs) as presented by (Horsman, 2020), the signification is paraphrased.

The OSAC framework postulates an entire range of possibilities for evaluative approaches. To the knowledge of the author, aside from the updated C-Scale (Casey, 2020) and the use of LR in combination with verbal scales (Casey et al., 2020a; Marquis et al., 2016), none of the proposed approaches have actually gained traction within the research community. To this day, examples where digital forensic results are properly evaluated are very rare. Instead, results are presented as factual, despite ample evidence that this should not be done.

2.2 Location-Related Digital Evidence

Definition 4. *A* **Location-Related Digital Trace** *is data generated by the operation of a mobile device as a function of its geographical location. (Casey et al., 2020a)*

In this work, it is proposed to further separate Location-Related Digital Evidence into Localisations and Location-Related Features, as these categories have a distinct behaviour regarding how they are evaluated. This separation is done to raise awareness of the fundamental difference regarding the information carried by those traces.

2.2.1 Localisation

Definition 5. *A* **Localisation** *is a location-related digital trace that results from a process attempting to determine the geographical location of the mobile device.*

Examples of localisation contain results from GPS, A-GPS or cell tower- and WiFi-triangulation. A localisation can be visualised through a coordinate (e.g. longitude and latitude) and has an associated accuracy. Consequently, it is important to consider accuracy, which can vary highly depending on the technology used. Table 2.4 gives an overview of accuracies of the most common localisation technologies.

Instead of looking at individual technologies, other researchers have looked at precision of localisation as a whole. Their studies and reported average precisions are presented in Table 2.5. In Rodriguez et al. (2018), the accuracy of Google timeline entries, the locations stored by Google based on the information it receives from an Android device linked to a Google account, were tested. The authors report an almost 50% rate of locations outside the indicated uncertainty radius for any localisation. (Rodriguez et al., 2018)

Technology	Accuracy	Source
GPS	5m	van Diggelen and Enge (2015)
WiFi (Signal Strength)	0.4m-40m	Kotaru et al. (2015)
WiFi (Fingerprinting)	5m-40m	Maghdid et al. (2016)
BLE	1m-10m	Faragher and Harle (2014)

Table 2.4: Accuracy of common localisation technologies. Omitted are Inertia sensors as no study on their accuracy was found.

Study	Sample size	Device Type	Precision
Syed et al. (2013)	5	Selected	7m-17m
PlaceIQ (2016)	150	No restrictions	30m
Merry and Bettinger (2019)	1	iPhone 6	7m-13m

Table 2.5: Reported precision in studies looking at localisation process in mobile devices as a whole.

It is important to note that the error on a localisation cannot always be simply modeled as a random error around the true location of the device. Systematic error may be introduced through effects such as shadowing and mirroring, where the direct view of a satellite is obstructed, but a reflective surface allows the device to communicate with the satellite anyway. As this reflected beam takes an indirect way to the device, its time of travel is longer than would be expected based purely on the location. This can introduce uncertainties, especially in urban regions where such topology is frequent (Kos et al., 2010). Merry and Bettinger (2019) have shown this effect in an extensive study with an iPhone 6 comprising in total 955 individual measurements to study the influence on accuracy of different parameters. Their results notably show a directional bias on some of the survey sites (Merry and Bettinger, 2019).

Errors in both the storing and the interpretation process of the data can lead to erroneous data. The most famous example of this is likely the so called «null-island», a non-existing place off the coast of Africa located at the origin of the global coordinate system. «Null-island» tends to be populated by localisations as a consequence of a tool interpreting or storing the absence of data (a NULL-Value) as the value 0, causing a localisation at the coordinates at 0,0 instead of no localisation at all (St. Onge, 2016). Other known effects on localisation can result from synchronisation delays or tampering with the internal clock. Through intentional manipulation, the localisation process can be completely bypassed and any location can be given to the device. This process called geo-spoofing is known to have been used by

players of augmented reality games such as Pokemon Go (Harber-Lamond, 2020). GPS relying on exchange of the device with the satellites, it is possible to induce external spoofing by overpowering the signal sent by the satellites by emitting a stronger signal in proximity (Eichelberger et al., 2020). It has been reported, that the Russian government may be using active geo-spoofing in order to stop Drone attacks on their president (Burgess, 2019).

2.2.2 Location-Related Feature

Definition 6. *A* **Location-Related Feature** *is a location-related digital trace that results as a by-product from a process whose primary function is not to determine the geographical location of the mobile device.*

As such, location-related features are all types of location related evidence that are not localisations. Examples of location-related features are frequent. They contain network connections, stored LAC in SIM cards, or multimedia content. Location-related features do not position the phone, but put the phone in an area with particular characteristics, for example, the area from which the device can connect to a certain cell tower. Consequently, this information can and should not be represented through coordinates with a related accuracy. Indeed the information "was within this area" cannot have an accuracy assigned.

Probably the most frequent type of location-related features are connections to cell towers. Even before mobile devices systematically recorded connections with them, they were used as evidence in court, as the information through which cell tower a call or SMS was routed is available to the phone service provider. These call data records, or CDR for short, indicate the location of said cell tower, giving evidence that the device was in the area in which this particular cell tower is accessible. Whilst in theory, the closest cell tower is expected to be the one to which a device connects, reality is much more complex. Especially in landscapes with an expansive and uninterrupted view, such as in close proximity to lakes, mobile devices are known to sometimes connect to far away cell towers. Measurements must therefore be made in the field using appropriate equipment to identify what towers serve a given location Tart et al. (2019); Jovanovic and Cummings (2022). Tart et al. (2012) have conducted a series of experiments where they have shown that obtaining complete data of serving cell towers is hard to obtain. In a series of measurements both with spot samples and location surveys, it is possible to miss serving cells for a given location. In their experiment, only area survey measurements in a 300m radius have reliably managed to detect all serving sites for the location they surveyed, however also picking

up a series of neighbouring cells as serving as well (Tart et al., 2012). A 2021 study by the same authors looked at whether these measurement devices are able to detect cells to which a series of controlled devices connected to. Depending on the measurement approach and the network provider, between 78% and 100% of cells connected to were detected (Tart et al., 2021). A developing question regarding cell site analysis is how to evaluate the recorded measurements when both proposed sites are served by the observed cell, as was the case in (Circuit Court of Albermarle County, 2015). The use of Bayesian logic is advocated in several publications (Tart et al., 2019; Tart, 2020; Tart et al., 2021; Casey et al., 2020a). So far, no full statistical model has been published, although first results from models in development have been presented (Bosma, 2022).

A very similar functionality as with cell towers may be applied to WiFi access points. Two main differences exist: First, the WiFi having a lower range, the area within which it is possible to connect to a WiFi access point is smaller, meaning that a lower uncertainty about the whereabouts of a device exists. Second, whilst cell tower infrastructure is generally owned and maintained by a limited number of enterprises that detain connection information in a centralised manner, WiFi infrastructure is often owned by a wide range of actors, from individuals to restaurants, enterprises or public organisations. Consequently, in addition to it being more difficult to obtain information from the network side, there is also generally no official register of the sites where the access points are located. Open source registers based on crowd sourced measurements exist, but are often incomplete and not particularly up to date. As with public cell tower databases, they do not show the effective site of the tower, but an estimated location based on measurements (Fu et al., 2012; Amundsen and Ovens, 2017).

Multimedia has for a long time been used as a matter of location identification. The use of visual media as a matter to show the presence of a person at a given place is frequent, both through the use of surveillance cameras or pictures taken on mobile devices, both by bystanders and persons involved in the matter. To the knowledge of the author, no scientific publication exists on this topic. A likely reason for this is that the location of the pictures is rarely contested and it would in many cases likely be possible to rule out most presented alternative locations. This approach has, however, gained traction in open source communities where potential locations are rarely known at the beginning or inaccessible to the investigator. Such approaches have been made possible by the surge of social media and the wide availability of open source geo-located information. Two prominent initiatives exploiting available online imagery to localise events are presented here for illustrative purpose. First, the "Trace an object"-initiative by Europol where

the public is asked to recognize an object, building or landscape outcropped from child abuse imagery profits from the internet by showing the images to a wide public. The underlying idea is that the localised availability of some objects allows to restrict the potential area in which the video was shot and therefore allows to be more specific in the search for potential suspects (EuroPol, 2017). Second, the open source initiative Bellingcat has used the wide availability of satellite images and verified localised imagery to localise videos and images from contested origin. Amongst other, they used buildings to localise an execution (Fiorella, 2020), followed Venezuelan politicians through Europe using social media posts and landscape imagery (Bellingcat Investigation Team, 2017) and localised ISIS training camps (Bellingcat Investigation Team, 2014). All without having ever been on site (Higgins, 2014).

As humans rely heavily on visuals, the use of location related features in audio recordings may be less intuitive to many. Nevertheless, such an approach is prevalent in popular media, for example in an episode of the crime series «NCIS» (Libman, 2006), or the third season of supernatural horror series «Stranger Things» (Levy, 2019). In reality, acoustic environment identification (AEI), started developing around 1980 in relation to forensic analysis of magnetic band recordings. Whilst the appearance of digital recordings initially complicated things due to low quality and compression, the rise of personalised mobile devices also made way to audio recordings being frequent traces found on mobile devices. Modern techniques in AEI primarily follow two paths: The first approach is based on the influence of the surroundings on the sounds recorded. In this approach, the surroundings themselves do not produce a sound, but reflect and change sound produced within them. Statistical analysis of reverberation has shown to provide good results for distinguishing small spaces, large spaces and outside recordings (Malik, 2013; Patil et al., 2019). Whilst this approach can be useful to exclude or categorize locations, it may only have limited use in the identification of a specific location. The second approach is based on elements in the surroundings of the recording device actively producing sounds. These often chaotic sounds can be modeled as a dynamic probabilistic process, which could be characteristic of a particular location. So far, such approaches have only been used to detect manipulation of audio (Ikram and Malik, 2010). Combined approaches creating a feature vector composed of both background noise and reverberation characteristics have been proposed and shown to be able to distinguish between locations with reasonable error rates in their respective experimental setting (Zhao and Malik, 2012, 2013; Moore et al., 2013). Whilst a potentially promising approach, AEI is to be considered a developing branch of digital audio forensics (Zakariah et al., 2018).

In some cases, location-related features may contain sufficient information so that further analysis can provide a result that has the same characteristics as a localisation. This has been done by experts to reconstruct the position of an object based on image, for example an unmarked grave based on the images taken on the day of the funeral 30 years prior to the exhumation (Pless et al., 2013). It is important to note that this analysis process does not change the nature of the traces it is based on. Similarly, the result of this analysis is not a trace, but a reconstruction based on the traces.

2.3 User Identification

Identifying the user of a device is a core challenge of system security. Inspired by classifications of authentication factors for access control (Dasgupta et al., 2017; Casey and Jaquet-Chiffelle, 2017), traces of identity on a mobile device can be classified in the following categories:

- Something the user *is*
- Something the user *does*
- Something the user *has*
- Something the user *knows*
- *Somewhere* the user is

Something the user *is*: Mobile devices contain user names, linked accounts, associated email-addresses and so on that may be close to the name or known nickname of someone (Casey and Turnbull, 2011). Stored contacts may give information about the persons relation, as for example the contact stored as «Mom» may actually be the persons mothers phone number.

Traces of physical characteristics can often be found in media recordings found on the device. Many characteristics that can be observed within images of the device's user are the subject matter of forensic disciplines predating mobile devices. Mentioned in literature are notably faces, finger- and palm-marks (or in the case of phone pictures, visible ridge skin patterns), ears, irises, scars, tattoos, as well as vein- and knuckle patterns. Whilst analysis and comparison is in many cases still based on manual comparison by an expert, automated systems for comparison are used in several domains (Champod and Tistarelli, 2017). Additionally, devices have started to incorporate physical biometry as active unlocking mechanisms and both

fingerprint detectors and facial recognition software are common features of modern mobile devices (Bhagavatula et al., 2015).

Something the user ***does*****:** How a user interacts with his device can be quite characteristic and allow for identification. The domain occupying itself with just that is called Behavioural Biometrics. Behavioural approaches of biometry can be classified in five categories based on the type of information that is collected about the individual to be identified (Yampolskiy and Govindaraju, 2008). Description of those categories and examples can be found in Table 2.6. It is important to note, that a specific technique may be part of multiple categories.

Category	Description	Examples
Authorship	Identification based on the characteristics of a piece of text, code or art created by the individual	Vocabulary, punctuation, painting style, coding style
Direct Human-Computer interaction (HCI)	Identification based on direct interaction with a digital device. May be software- or input device-based.	Keyboard & mouse dynamics (input device), command line lexicon (software)
Indirect HCI	Identification based on traces resulting from user interaction.	Audit logs, network traffic, system calls
Motor-Skills	Identification based on muscle movements	Walking pattern, vocal recognition, lip movement
Purely Behavioural	Identification based on behaviour not directly concentrating on measurements of body parts.	Calling behaviour, driving style, credit card usage

Table 2.6: Categories of behavioural biometric approaches according to (Yampolskiy and Govindaraju, 2008)

Something the user ***has*****:** Many mobile devices contain SIM-cards, that, in many countries, may only be purchased with identification (Casey and Turnbull, 2011). Additionally, as with physical biometry, objects, such as clothing, the user possesses may be seen in imagery found on the device, and it may be possible to uniquely identify those objects (Jaha and Nixon,

2016). As this type of identification generates an additional indirection for identification, it is not considered in this work.

Something the user ***knows*****:** Something the user knows may be the PIN, password or pattern used to unlock the phone. Similarly, if the user uses the device to access a password protected service, types in a phone number or makes a note of something only a limited number of people know, this may allow some degree of identification.

Somewhere **the user is:** Places where a person frequently goes, such as his or her home address, place of work or preferred restaurant, pub or cinema may be found on a device (Casey and Turnbull, 2011) and allow to identify a person. Several authors have been interested in using location as a means of identifying the user of a device or account as the person using another device / account for which the user is known (Galbraith et al., 2020; Bosma et al., 2020; De Bie, 2022). As this work is interested in location-focused propositions, this category of identity-related traces is not further considered, as it may be cause for inadvertent co-dependence of traces.

Chapter 3

Mathematical Framework

As described in Chapter 1, at the core of this work is a mathematical framework for combining uncertainties from different phases in the reasoning process to produce an overall LR on the position of the person of interest. This chapter is structured as follows: The Bayesian network used as a basis for the framework is constructed and explained in Section 3.1. Then in Section 3.2 it is demonstrated mathematically that the Bayes Net behaves as expected in extreme cases. Finally, Section 3.3 shows the influence of variables that cannot be easily informed by external parameters on the final result.

3.1 Creation of the Bayesian Framework

A Bayesian Network is created to allow for a mathematical combination of the results. The network is structured in a way that contains a propositions-node for each step explained in Section 1.2 except the preliminary phase. As noted there, this authentication phase will be considered to have already been completed with an affirmative conclusion for each question and each piece of evidence. Consequentially, no nodes for it appears in the Bayes Net[1]. The network is constructed from the proposition nodes outwards, adding evidence, and supplementary nodes to evaluate this evidence, where possible. Nodes deriving from one proposition node should be only interdependent with nodes deriving from the same proposition node. That way, the model becomes independent of the type of evidence used for reasoning. The scenarios studied and presented in this work are chosen in a way to ensure this independence. Whether it holds true for other cases should be reconsidered for each pairing

[1] If this assumption were not made, the Bayes Net would have to be configured in a way that the evidence becomes irrelevant if a piece of evidence fails one of the three preliminary tests.

of evidence.

The general structure of the model is shown without any evidence-nodes in Figure 3.1. This is an extension of the model presented in (Biedermann and Vuille, 2016) without the H node that denotes whether the accused is the perpetrator or not. The P node of Biedermann and Vuille's network is represented as the node $LocP$ and the node M as $LocD$. E is generalized to be any piece of evidence relating to the location of the device and shown in Figure 3.2 as E_1. The potential states of the respective nodes are extended to take into account a higher variability of scenarios. All propositions are shown in Table 3.1. Depending on the case at hand, it is likely that some states are ruled out by the parties. Under such circumstances, the excluded states can either be removed from the network or their probability can be set to zero. Due to the restriction of this work, the nodes about locality are restricted to two alternatives.

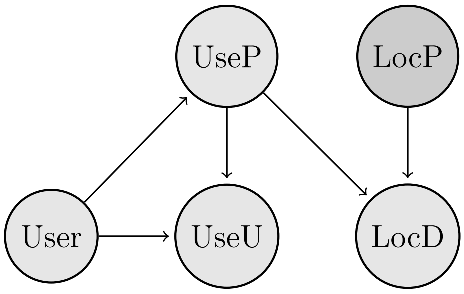


Figure 3.1: Bayesian Network containing the Proposition-Nodes of the phases cited in the previous section. The possible propositions are listed in Table 3.1

The Propositions of the node «LocP» are the ones in which the expert is interested in for the scope of the expertise. The nodes «UseU» and «LocD», as well as some states of «UseP», are in large part the result of logical combination of prior probabilities. The probability table for «UseP» is shown in Table 3.2. If no one is using the device in general (P_3 is true in node «User»), then no one is using the device at time t either (P_3 is true in node «UseP» as well). The other probabilities are not defined as such, but it is to be expected that α is larger than β and δ is larger than γ, as the probability of someone being the user at a given moment should be increased by the fact that this person is the general user of the device. The influence of the values α, β, γ and δ is studied in Section 3.3.

The probability table for node «UseU» is shown in Table 3.3. For most of the fields, 1 and 0 values can be inserted through logical comparison. Whenever either «UseP» or «User» is in state 3, corresponding to to «no

Node	Propositions / Node States
User	$User_1$: Person A is the general user of the device. $User_2$: Someone else is the general user of the device. $User_3$: No one is using the device.
UseU	$UseU_1$: The general user of the device is the user at time t. $UseU_2$: At time t, there is another user than the general user. $UseU_3$: No one is using the device at time t.
UseP	$UseP_1$: Person A is using the device at time t. $UseP_2$: Someone else is using the device at time t. $UseP_3$: No one is using the device at time t.
LocD	$LocD_1$: The device was located at location X at time t. $LocD_2$: The device was located at location Y at time t.
LocP	$LocP_1$: The person was located at location X at time t. $LocP_2$: The person was located at location Y at time t.

Table 3.1: States of all nodes in Figure 3.1

User		P_1	P_2	P_3
	P_1	α	γ	0
UseP	P_2	β	δ	0
	P_3	$1-\alpha-\beta$	$1-\gamma-\delta$	1

Table 3.2: Probability table of node UseP.

one using the devices», «UseU» also takes on state 3, also representative of no one using the device. If A is the general user of the device ($User_1$) and is also using the device at time t ($UseP_1$), then the general user is the one using the device at time t ($UseU_1$). And if the state of «User» and «UseP» designate a different person currently holding the device and being the general user, the state of «UseU» will always be $UseU_2$, signifying a different person than the general user is currently using the device. Only two fields contain a value different from 1 or 0. Indeed, the probability of the general user using the device if the general user is not Person A and someone different from Person A is currently using the device, is not evident. The influence of θ will be studied in Section 3.3, although, given the removed position of the «UseU»-node, the influence of θ is expected to be minor.

The probability table for «LocD» is shown in Table 3.4. Here, the node is purely logical because if the state of «UseP» is P_1, in other words, if the Person A is using the device at time t, the person location is equal to the device location. Indeed, if any other person is using the device, or if no one is using it, the position of the device cannot actually be relevant to determine

UseP		P_1			P_2			P_3		
User		P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
	P_1	1	0	0	0	θ	0	0	0	0
UseU	P_2	0	1	0	1	$1-\theta$	0	0	0	0
	P_3	0	0	1	0	0	1	1	1	1

Table 3.3: Probability table for node UseU.

any information about the position of the person. Consequently, the table is filled with equal probabilities of $^1/_2$ for all states in the remaining cases.

UseP		P_1		P_2		P_3	
LocP		P_1	P_2	P_1	P_2	P_1	P_2
LocD	P_1	1	0	$^1/_2$	$^1/_2$	$^1/_2$	$^1/_2$
	P_2	0	1	$^1/_2$	$^1/_2$	$^1/_2$	$^1/_2$

Table 3.4: Probability table for node LocD

For each node where direct evidence exists, nodes are added for the evidence. These nodes are derived from the propositions, as the evidence observed is a consequence of the action that took place. Figure 3.2 shows the network with added evidence nodes for all propositions where it is considered that direct evidence is possible. The formula for the LR of this Bayes Net is provided in Annex A. In the given example, the nodes represent the totality of evidence observed in relation to the respective proposition and are considered to be independent from each other. Whether or not this assumption is justified will depend on the considered evidence and should be verified in each case individually. Case examples in this work will be chosen in a way that the assumption can be reasonably justified, as co-dependence of evidence would render the problem exponentially more complex. Figure 3.2 considers evidence nodes as black boxes. The expert has no knowledge about the internal workings that influences the outcome of the evidence. The probability is purely derived from experiments where the situation was simulated and then the outcome was observed. If knowledge about the inner workings of the process is available, the network can be adapted respectively. Generally, studies that provide data for systems where the internal workings are easier to conduct and, therefore, are preferable to systems where this is not the case.

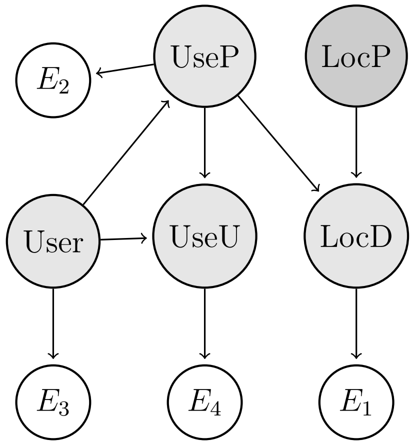


Figure 3.2: Bayesian Network containing the Proposition-Nodes and Evidence-Nodes.

3.2 Demonstration of Expected Behaviour in Extreme Cases

The Bayes Net constructed above has some extreme cases presented in Section 1.3 where we know the expected behaviour. Basically, if some information is categorically known, we can input them into the Bayes Net and see whether the network behaves as expected. In the following, formulaic demonstrations are conducted to show the following statements are true:

1. If Person A is known to be in the possession of the device at time t, the overall LR becomes equal to the LR at device level.
2. If Person A is known to not be in possession of the device at time t, the overall LR becomes 1.
3. If no person has the device at time t, the overall LR becomes 1.

This section shows fundamental validity of the framework: In cases where we know how the network should behave, we know that there is no fundamental flaw with the network.

3.2.1 Notation

The notation used in this work is as follows:

A_n	designates the state n of node A, e.g. $LocP_1$ indicates state 1 in node $LocP$.
E_n	designates the observation n.
$Pr(A)$	designates the probability of A.
$Pr(A \mid B)$	designates the probability of A given that B is known to be true.

3.2.2 Person A is known to be in possession of the device at time t

In the case where it is categorically known that Person A was in possession of the device at time t, the location of the device becomes automatically the location of the person. Following the same logic, uncertainties regarding the location of the device must be the same for the location of the person. In other words, the overall LR (on person-level), written as follows :

$$LR = \frac{Pr(E_1, E_2, E_3, E_4 \mid LocP_1)}{Pr(E_1, E_2, E_3, E_4 \mid LocP_2)} \tag{3.1}$$

is equal to the LR at device level, written as :

$$LR = \frac{Pr(E_1, E_2, E_3, E_4 \mid LocD_1)}{Pr(E_1, E_2, E_3, E_4 \mid LocD_2)} \tag{3.2}$$

It is shown in this section that this is indeed the case. To do so, a property of Bayes Nets is used called "screen off"-effect. This property states, that if the state of a node is known, the state of this nodes parent-node will not have any further influence on child-nodes of this node, as long as there are no other connections between the parent-nodes and the child-node (Taroni and Aitken, 2006). In the here specified situation, the state of the node $UseP$ is known, causing a screen-off effect between the identity-focused nodes on the left of the network and the location-focused nodes on the right. To obtain a formula for the LR in a Bayes Net where the state of node $UseP$ is known, the identity-focused nodes as well as any evidence nodes declined from $UseP$ loose their relevance. This leaves the Bayes Net shown in Figure 3.3.

As can be seen, all evidence-nodes but E_1 have been removed. In the following, the formula for this reduced Bayes Net is developed. This will be done for a unfixed state of UseP, so that the formula can be reused at a later stage.

$$LR = \frac{Pr(E_1 \mid LocP_1)}{Pr(E_1 \mid LocP_2)} \tag{3.3}$$

To be able to properly represent the formulas, numerator and denominator will initially be treated separately.

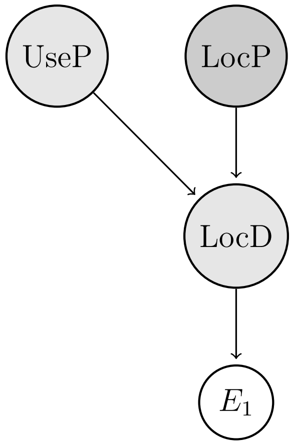


Figure 3.3: Reduced Bayesian Network for the case where the possession of the device is known.

Numerator: The numerator is equal to:

$$\begin{aligned} & Pr(E_1 \mid LocD_1; LocP_1) \cdot Pr(LocD_1 \mid LocP_1) \\ & \quad + Pr(E_1 \mid LocD_2; LocP_1) \cdot Pr(LocD_2 \mid LocP_1) \end{aligned} \tag{3.4}$$

Given that the E_1-node has only *LocD* as a parent node, as a consequence of the screen-off effect, if a probability for E_1 is conditioned by a *LocD*-state, all other conditions become irrelevant and can be removed.

$$Pr(E_1 \mid LocD_1) \cdot Pr(LocD_1 \mid LocP_1) + Pr(E_1 \mid LocD_2) \cdot Pr(LocD_2 \mid LocP_1) \tag{3.5}$$

Including the UseP node

$$\begin{aligned} & Pr(E_1 \mid LocD_1) \cdot [Pr(LocD_1 \mid LocP_1; UseP_1) \\ & \quad \cdot Pr(UseP_1) + Pr(LocD_1 \mid LocP_1; UseP_2) \cdot Pr(UseP_2) \\ & \quad + Pr(LocD_1 \mid LocP_1; UseP_3) \cdot Pr(UseP_3)] \\ & \quad + Pr(E_1 \mid LocD_2) \cdot [Pr(LocD_2 \mid LocP_1; UseP_1) \\ & \quad \cdot Pr(UseP_1) + Pr(LocD_2 \mid LocP_1; UseP_2) \cdot Pr(UseP_2) \\ & \quad + Pr(LocD_2 \mid LocP_1; UseP_3) \cdot Pr(UseP_3)] \end{aligned} \tag{3.6}$$

This expression now only contains values that are in the probability tables of the Bayes Nets nodes. By substituting with the values from the *LocD*-node (cf. Table 3.4), the expression becomes:

$$\begin{aligned} & Pr(E_1 \mid LocD_1) \cdot [1 \cdot Pr(UseP_1) + 1/2 \cdot Pr(UseP_2) + 1/2 \cdot Pr(UseP_3)] \\ & \quad + Pr(E_1 \mid LocD_2) \\ & \quad \cdot [0 \cdot Pr(UseP_1) + 1/2 \cdot Pr(UseP_2) + 1/2 \cdot Pr(UseP_3)] \end{aligned} \tag{3.7}$$

which is equal to

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [Pr(UseP_1) + {}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \\ &\quad + Pr(E_1 \mid LocD_2) \cdot [{}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \end{aligned} \tag{3.8}$$

If the device is known to be in possession of Person A at time t, then $Pr(UseP_1) = 1$, $Pr(UseP_2) = 0$ and $Pr(UseP_3) = 0$. The numerator therefore becomes:

$$Pr(E_1 \mid LocD_1) \cdot [1 + {}^1\!/_2 \cdot 0 + {}^1\!/_2 \cdot 0] + Pr(E_1 \mid LocD_2) \cdot [{}^1\!/_2 \cdot 0 + {}^1\!/_2 \cdot 0] \tag{3.9}$$

which reduces to

$$Pr(E_1 \mid LocD_1) \tag{3.10}$$

Denominator: The same reasoning is followed as with the numerator. The denominator is equal to :

$$Pr(E_1 \mid LocD_1) \cdot Pr(LocD_1 \mid LocP_2) + Pr(E_1 \mid LocD_2) \cdot Pr(LocD_2 \mid LocP_2) \tag{3.11}$$

Inluding $UseP$:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [Pr(LocD_1 \mid LocP_2; UseP_1) \\ &\quad \cdot Pr(UseP_1) + Pr(LocD_1 \mid LocP_2; UseP_2) \cdot Pr(UseP_2) \\ &\quad + Pr(LocD_1 \mid LocP_2; UseP_3) \cdot Pr(UseP_3)] \\ &\quad + Pr(E_1 \mid LocD_2) \cdot [Pr(LocD_2 \mid LocP_2; UseP_1) \\ &\quad \cdot Pr(UseP_1) + Pr(LocD_2 \mid LocP_2; UseP_2) \cdot Pr(UseP_2) \\ &\quad + Pr(LocD_2 \mid LocP_2; UseP_3) \cdot Pr(UseP_3)] \end{aligned} \tag{3.12}$$

Substituting from Table 3.4:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [0 \cdot Pr(UseP_1) + {}^1\!/_2 \cdot Pr(UseP_2) + {}^1\!/_2 \cdot Pr(UseP_3)] \\ &\quad + Pr(E_1 \mid LocD_2) \\ &\quad \cdot [1 \cdot Pr(UseP_1) + {}^1\!/_2 \cdot Pr(UseP_2) + {}^1\!/_2 \cdot Pr(UseP_3)] \end{aligned} \tag{3.13}$$

which is equal to

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [{}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \\ &\quad + Pr(E_1 \mid LocD_2) \cdot [Pr(UseP_1) + {}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \end{aligned} \tag{3.14}$$

Knowing that the device was held by person A at time t:

$$Pr(E_1 \mid LocD_1) \cdot [{}^1\!/_2 \cdot 0 + {}^1\!/_2 \cdot 0] + Pr(E_1 \mid LocD_2) \cdot [1 + {}^1\!/_2 \cdot 0 + {}^1\!/_2 \cdot 0] \tag{3.15}$$

which reduces to

$$Pr(E_1 \mid LocD_2) \tag{3.16}$$

Combining into a fraction: Bringing together numerator and denominator, the LR becomes:

$$LR = \frac{Pr(E_1 \mid LocD_1)}{Pr(E_1 \mid LocD_2)} \tag{3.17}$$

Which is equal to the LR at device-level. *Q.E.D.*

3.2.3 Person A is known to not be in possession of the device at time t

If there is categorical knowledge that the person was not in possession of the device at time t, then the evidence of the devices location cannot give any relevant evidence on the whereabouts of this person. Non-pertinent evidence is mathematically expressed by an LR of 1. In the following it is shown, that if the person is known to not be in the possession of the device, the LR always becomes 1.

Numerator: Starting from Formula 3.8, the numerator is:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [Pr(UseP_1) + {}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \\ &+ Pr(E_1 \mid LocD_2) \cdot [{}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \end{aligned} \tag{3.18}$$

This time, the person is known to not be in possession of the device at time t. So, $Pr(UseP_1) = 0$, $Pr(UseP_2) = 1$ and $Pr(UseP_3) = 0$. The numerator therefore becomes:

$$Pr(E_1 \mid LocD_1) \cdot [0 + {}^1\!/_2 \cdot 1 + {}^1\!/_2 \cdot 0] + Pr(E_1 \mid LocD_2) \cdot [{}^1\!/_2 \cdot 1 + {}^1\!/_2 \cdot 0] \tag{3.19}$$

Which reduces to:

$${}^1\!/_2 \cdot Pr(E_1 \mid LocD_1) + {}^1\!/_2 \cdot Pr(E_1 \mid LocD_2) \tag{3.20}$$

Denominator: Starting from Formula 3.14, the denominator is:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [{}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \\ &+ Pr(E_1 \mid LocD_2) \cdot [Pr(UseP_1) + {}^1\!/_2 Pr(UseP_2) + {}^1\!/_2 Pr(UseP_3)] \end{aligned} \tag{3.21}$$

Following the same logic as with the numerator:

$$Pr(E_1 \mid LocD_1) \cdot [{}^1\!/_2 \cdot 1 + {}^1\!/_2 \cdot 0] + Pr(E_1 \mid LocD_2) \cdot [0 + {}^1\!/_2 \cdot 1 + {}^1\!/_2 \cdot 0] \tag{3.22}$$

Which reduces to:

$${}^1\!/_2 \cdot Pr(E_1 \mid LocD_1) + {}^1\!/_2 \cdot Pr(E_1 \mid LocD_2) \tag{3.23}$$

Combining into a fraction: Bringing together numerator and denominator, the LR becomes:

$$\begin{aligned} LR &= \frac{1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2)}{1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2)} \\ &= 1 \end{aligned} \tag{3.24}$$

Q.E.D

3.2.4 No person is in possession of the device at time t

If there is categorical knowledge that no person was in possession of the device at time t, then the evidence of the devices location cannot give any relevant evidence on the whereabouts of any person. Following the same logic as in Section 3.2.3 it is shown, that if the device is known to not be in the possession of anyone at time t, the LR always becomes 1.

Numerator: Starting from Formula 3.8, the numerator is:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [Pr(UseP_1) + 1/2 Pr(UseP_2) + 1/2 Pr(UseP_3)] \\ &+ Pr(E_1 \mid LocD_2) \cdot [1/2 Pr(UseP_2) + 1/2 Pr(UseP_3)] \end{aligned} \tag{3.25}$$

This time, the person is known to not be in possession of the device at time t. So, $Pr(UseP_1) = 0$, $Pr(UseP_2) = 0$ and $Pr(UseP_3) = 1$. The numerator therefore becomes:

$$Pr(E_1 \mid LocD_1) \cdot [0 + 1/2 \cdot 0 + 1/2 \cdot 1] + Pr(E_1 \mid LocD_2) \cdot [1/2 \cdot 0 + 1/2 \cdot 1] \tag{3.26}$$

Which reduces to:

$$1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2) \tag{3.27}$$

Denominator: Starting from Formula 3.14, the denominator is:

$$\begin{aligned} &Pr(E_1 \mid LocD_1) \cdot [1/2 Pr(UseP_2) + 1/2 Pr(UseP_3)] \\ &+ Pr(E_1 \mid LocD_2) \cdot [Pr(UseP_1) + 1/2 Pr(UseP_2) + 1/2 Pr(UseP_3)] \end{aligned} \tag{3.28}$$

Following the same logic as with the numerator:

$$Pr(E_1 \mid LocD_1) \cdot [1/2 \cdot 0 + 1/2 \cdot 1] + Pr(E_1 \mid LocD_2) \cdot [0 + 1/2 \cdot 0 + 1/2 \cdot 1] \tag{3.29}$$

Which reduces to:

$$1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2) \tag{3.30}$$

Combining into a fraction: Bringing together numerator and denominator, the LR becomes:

$$
\begin{aligned}
LR &= \frac{1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2)}{1/2 \cdot Pr(E_1 \mid LocD_1) + 1/2 \cdot Pr(E_1 \mid LocD_2)} \\
&= 1
\end{aligned}
\tag{3.31}
$$

Q.E.D

3.3 Behaviour of the Bayes Net

To understand the functionality of the Bayes Net and to study the impact of the various parameters in the network, Bayesian Networks were modeled in the decision-making software Hugin version 8.3 (Hugin Expert). Using the R-library RHugin[2] (Konis and Moharil, 2008), these Networks were imported into R where the impact of the parameters was studied by varying the values of interest from 0 to 1 and compiling the output for each setting. The LR for each setting is observed and plotted.

As shown in Section 3.2.2, the person-level LR becomes the device-level LR if it is categorically known that a specific person was in possession of the device at the moment of interest. As this is the situation where the least uncertainty exists, this situation corresponds to the LR the furthest away of 1. In the other extreme, the LR becomes 1 if it is categorically excluded that the device was in possession of the person of interest, as the location-related evidence recovered from the device becomes non-pertinent for the question of the position of this person.

To illustrate the behaviour in between those extremes, a simplified Bayes Net was created where only the node «UseP» in addition to the location-related nodes was added (cf. Figure 3.4) and the LR was mapped as a function of the probability of Person A being in possession of the device. This simulations allows to understand the behaviour as a function of the degree of certainty in the fact that Person A was the user of the device at time t. The result is shown in Figure 3.5. As can be seen, for the interval 0 to about 0.9, the person-level LR remains quite low, only to rise sharply afterwards until the device-level LR is attained. This shows that a high posterior believe in usage has to be obtained to reach a meaningful person-level LR.

[2]RHugin has not been updated since 2017 and is not supported for current versions of R. These simulations[3] were run on a computer that has not been updated since 2017. This is obviously not a sustainable approach and effort into developing a stable framework allowing to do simple manipulations in Python or R should be considered in the future.

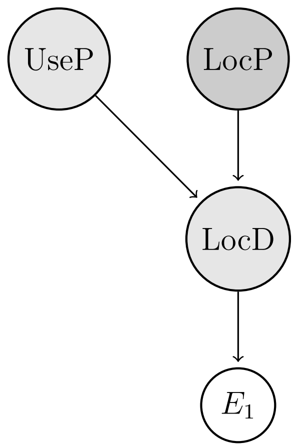


Figure 3.4: Simplified network to study the influence of $UseP$.

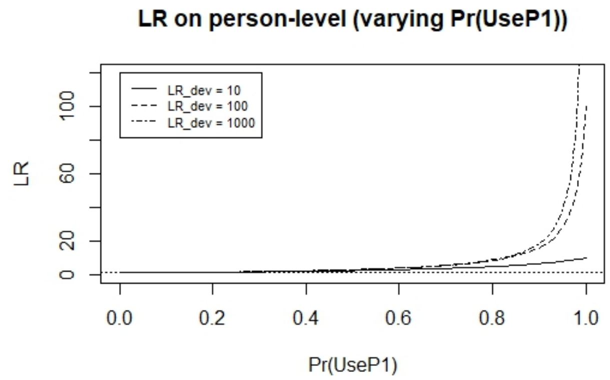


Figure 3.5: Person-level LR as a function of $Pr(UseP_1)$ from the Bayes Net as shown in Figure 3.4. The impact is shown for device-level LR of 10, 100 and 1'000.

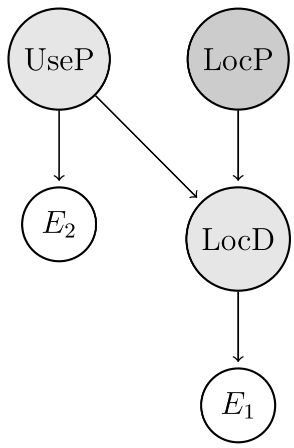


Figure 3.6: Bayesian Network with direct evidence of usage: a second evidence node is added to «UseP» to study the influence of UseP-priors.

The posterior probability of $UseP_1$ is constituted of the prior probability of $UseP_1$ and the weight of the evidence provided in favour of $UseP_1$. In cases where a conclusion on the person-level is aimed to be achieved, it should be aimed for the evidence being the dominant factor in the posterior probability of $UseP_1$. To study the impact of the prior probabilities, the evidence node «E_2» is added to the Bayes Net resulting in the network shown in Figure 3.6. It's impact is shown in Figure 3.7. As can be seen, the priors only have a major impact if the supporting evidence is not particularly strong. This leads to the general recommendation to only express opinions on the person-level in situations like this, if evidence of an overall LR in favour of possession of 100 or higher is present. In this case, the distribution is reasonably stable for justifiable range of priors in UseP.

In cases where indirect evidence of usage is considered, four variables that cannot be clearly fixed have to be considered in the «UseP»-node. The Bayes Net for this situation is shown in Figure 3.8. These variables were designated as follows in Table 3.2:

- α for $Pr(UseP_1 \mid User_1)$
- β for $Pr(UseP_2 \mid User_1)$
- γ for $Pr(UseP_1 \mid User_2)$
- δ for $Pr(UseP_2 \mid User_2)$

α is inherently linked to β, as their sum cannot surpass 1. The same is true for γ and δ. The analysis for those four variables is done in two series:

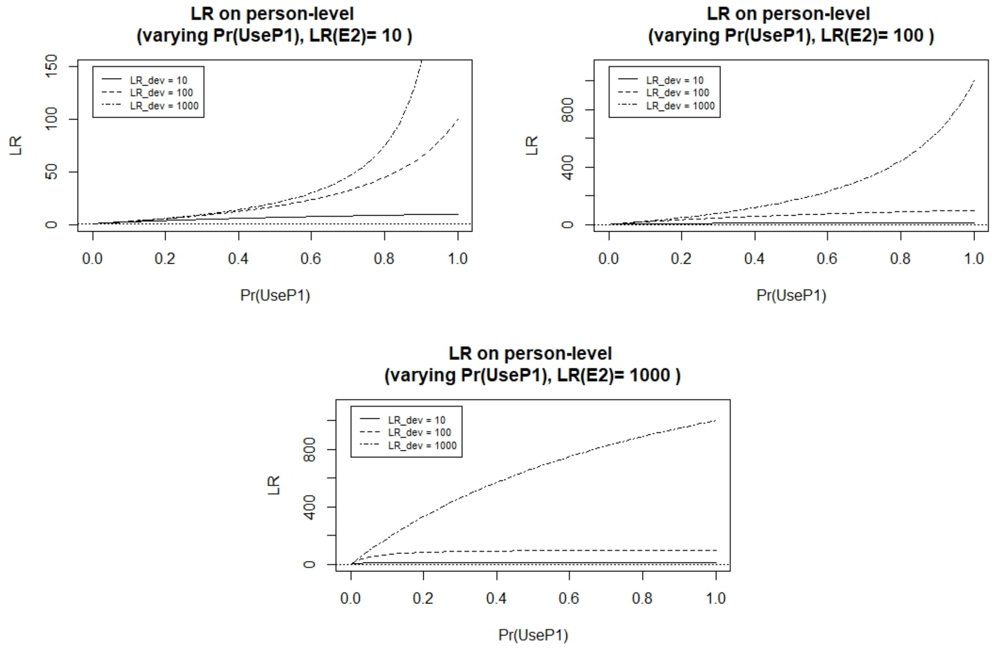


Figure 3.7: Simulations of the influence of $Pr(UseP_1)$ on the subject-level LR ($LocP$-Node) for the Bayes Net shown in Figure 3.6. The simulation is run for varying levels of device-level LR and LR on the User of 10 (top left), 100 (top right) and 1000 (bottom).

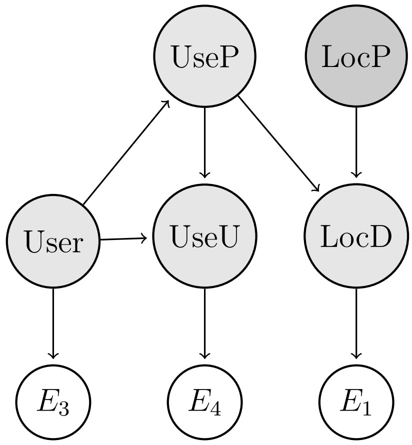


Figure 3.8: Bayesian Network with indirect evidence of usage.

First, α is studied with several fixed values for γ. β is modeled as a fraction of (1-α) with four different values. Second, the analysis is inverted, with varied γ, δ as a fraction of (1 - γ) and multiple fixed values for α.

The results can be seen in Figure 3.9 for α and Figure 3.10 for γ. As can be seen, there is little influence of α above 0.3. As it is to be expected that a person is more likely to have their own phone than someone else, this is a range in which the value is expected to be. The value of β mostly influences the form of the distribution. The lower β becomes in relation to α, the quicker the distribution plateaus.

For γ, the value is expected to be in the lower range, as it is the probability of Person A having the device is expected to be less likely. As can be seen in Figure 3.10, this is not too much of an issue, as the value of γ has only minor impact on the overall LR. The impact of δ is negligible.

Simulations on θ and $Pr(User_1)$ show that these variables have close to no influence on the overall LR.

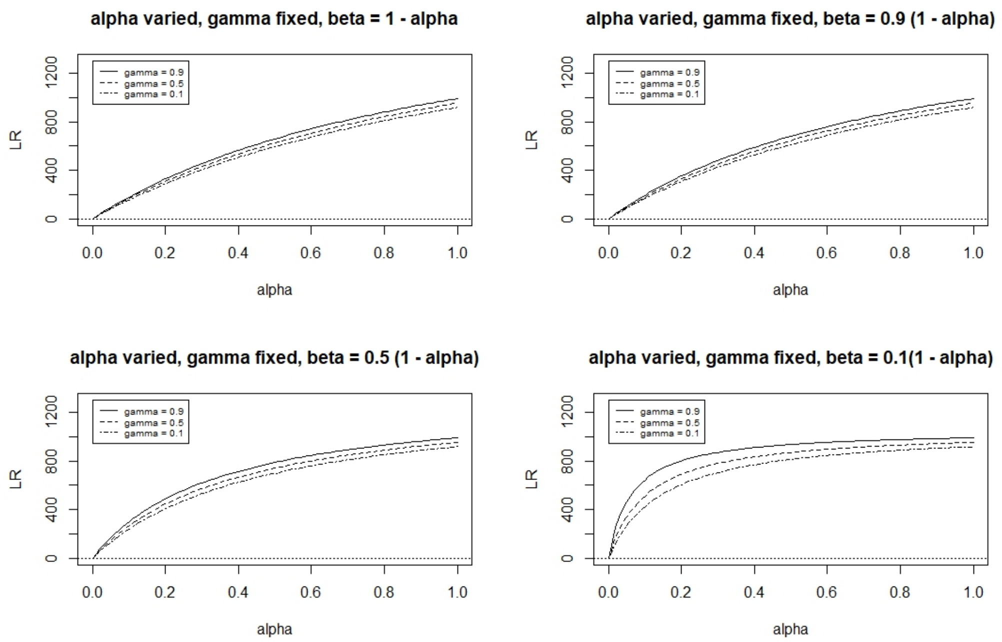


Figure 3.9: Simulation of person-level LR (*LocP*-node) as a function of α

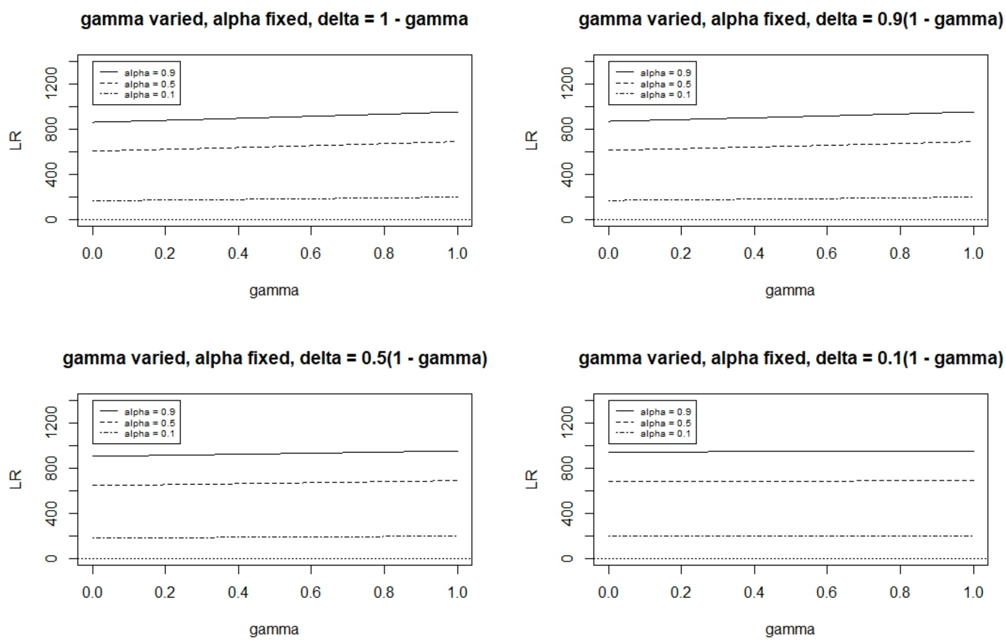


Figure 3.10: Simulation of person-level LR (*LocP*-node) as a function of γ

Chapter 4

Scenario 1

The aim of this chapter is to illustrate the use of the model in a simplified manner: a single piece of location-related evidence is evaluated in light of propositions on the level of the device. This setting can be applied when phone possession is not contested or if it is established through means outside of the experts domain of competence, such as witness testimony. As an example, a single piece of cell site evidence, the connection established when placing a phone call, is used as a scenario that is commonly encountered by law enforcement agencies.

4.1 Description of the Scenario

The aim of this scenario is to illustrate the reasoning and use of an LR at the device level. The question of interest is from which of two specific positions was a phone call made. This problem may arise in situations such as described below:

> A suspect is accused of having committed a crime at a given address. The accused contests having been at the crime scene, claiming that he was at home the entire time whilst the crime took place. Call data records (CDR) are presented as evidence. During the time frame of the crime, a phone call was recorded coming from the accused's mobile phone. The person with whom he was talking confirms his identity.

In such a situation, an expert tasked with evaluating the observed evidence would have to assess whether the evidence is more likely to be observed if the phone was at the crime scene or the accused's house. The witness testimony should be considered outside of the domain of competence of the expert. The following pair of device-level propositions are likely to be considered:

P_1:	The mobile device was at Location X (the crime scene) at the time of the phone call.
P_2:	The mobile device was at Location Y (the accused's home) at the time of the phone call.

In most real life situations, the first question to consider is whether the cell site in question is actually accessible from both locations. If this is not the case for one of the sites, it can be ruled out and an evaluative assessment may at most be made about the chance of an error in either the evidentiary or the reference data. This situation is, however, outside the scope of this research as device and measurement errors are considered to be excluded from this work as stated in Section 1.2. To illustrate a probabilistic assessment based on data assumed to be free of error, locations are chosen that are both covered by the same cell tower as the tower observed in evidence.

4.2 Theoretical Background of Cell Tower based Localisation

Mobile device communication has a core issue it needs to overcome: The handheld devices do not have the antenna, nor the battery life, to send out signals strong enough to reach an arbitrarily distant receiver it aims to communicate with. The solution to this problem are cell towers. A network of antennas spread out over the countryside allowing the devices to connect with a world spanning communication network and with other end devices, in turn also connected to a cell tower themselves (cf. Figure 4.1). Generally, cell towers are provisioned in a way that they all serve more or less the same quantity of devices. As devices are more prevalent in more populated places, cell tower ranges can vary significantly from several kilometers in the countryside to just a block in cities. Heavily trafficked areas, such as airports, metro lines and shopping malls, may even have their own cell site covering just the interior of the building (Hoy, 2015). Quickly, the forensic value of these connections became evident. CDR kept by network providers both for billing and maintenance purposes, contain details about the cell tower that a phone call was routed through, and can be used to place a device in a particular region when the phone call was made. With the evolution of smartphones and mobile data, the temporal granularity has increased as there is an almost constant exchange of data, and network providers retain information about the region a device is operating in effectively at any given moment.

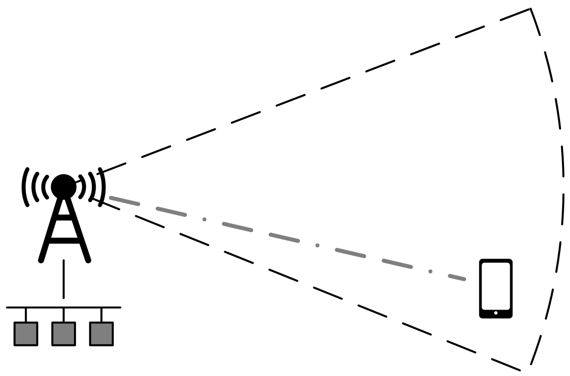


Figure 4.1: Illustration of the functionality of mobile networks. Within the coverage area of a cell tower, mobile devices can connect to this cell tower through a radio signal. The cell tower is in turn connected to a network allowing the end device to communicate with any other device connected to this network.

The evident utility of CDR in investigative contexts and the equally evident risk for abuse of this data has required legislators all over the world to formulate laws regulating the storage and access to this data (ISDC, 2013)[1] In Switzerland, service providers are required by law to keep these records for six months, after which point the data is to be erased. Law enforcement agencies can gain access with a warrant through an automated platform, maintained by the Post and Telecommunication Surveillance Service (PTSS), an independent unit within the federal department of justice and police. Additionally, the PTSS is responsible for the proper implementation of postal and telecommunication metrics, and publishes yearly statistics on the use of those metrics by the different law enforcement agencies within Switzerland (SPTA, 2016; PTSS, 2021). Figure 4.2 visualises the development of requests for CDR, called retroactive surveillance measures, by LEA from 2018 to 2021. The requests decrease from year to year, which could be explained by a increasing understanding of investigators in which cases there is an interest in requesting CDR.

[1]These laws have also been frequently criticised and challenged. For example, in April of 2022, the Court of Justice of the European Union found the Irish Law to violate European law which prohibits indiscriminate data retention and access without independent safeguard(CJEU, 2022).

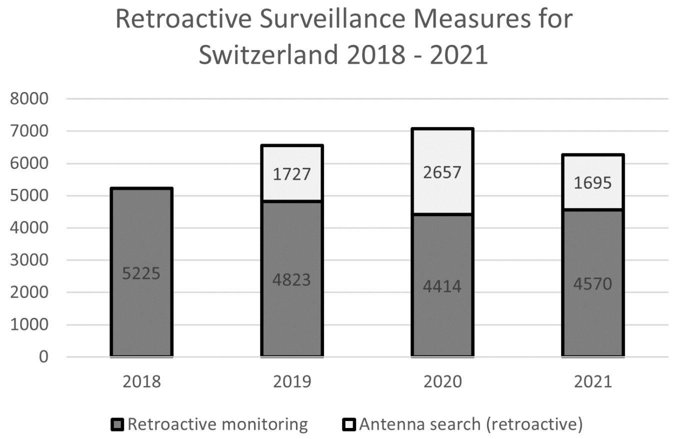


Figure 4.2: Use statistics of retroactive telecommunication surveillance in Switzerland from 2018 to 2021. Source of Data: (PTSS, 2021). For 2018, retroactive monitoring and antenna searches were not yet indicated separately. Data prior to 2018 is available, however not comparable with the shown quantities as the counting procedure was adapted in between.

4.2.1 Uncertainties in CDR Data Analysis

For CDR to be of use, one has to know which cell towers are accessible at a given site. The process of obtaining this information through measurements with specialised equipment is called cell site surveying. A multitude of approaches exist, from spot measurements, where connectivity at a singular location is measured, to cell coverage surveys, where the entire coverage of an antenna is mapped out. From a measurement action point of view, «Connected Mode»- and «Idle Mode»-surveys are distinguished. In Connected Mode, mobile devices with SIM cards are used to observe what connections are established. This approach is generally quite inefficient for obtaining an understanding of all available towers in a region since devices will only connect to towers from the operator they are subscribed to and might not be equipped to support all technologies and frequencies. Additionally, even if the pertinent technology and the operator are known in an investigation, an active device may not connect to an available antenna, when another is far more dominant in the vicinity. In Idle Mode, the measuring device does not interact with the cell towers, but measures the signal of all antennas visible from the point of measurement. This approach is generally more efficient because a single measurement can give an initial indication of accessible cell

towers[2] (Hoy, 2015). A major limitation of this approach is that there is currently no way to predict based on these measurements whether a device will actually connect to an observed cell tower. There seems to be very little awareness of this fact among practitioners. Indeed, some practitioners stipulate that a mobile device will always connect to the strongest available cell tower, due to physical laws of signal propagation, generally the closest one (Griffiths and Hoy, 2018). It is quite easy to demonstrate that this is not the case. Indeed, if it were, the probabilistic evaluation of cell tower evidence as presented in this chapter would, some fringe cases reserved, be completely without use, as anyone would be able to categorically predict the cell tower a device connects to for a given location. As can be seen from the data presented in this chapter, mobile devices do not behave in such a deterministic manner. Even in publications that acknowledge the possibility of devices not connecting the best serving cell, the frequency of this happening is considered to be quite low. Jovanovic and Cummings (2022) indicate a frequency of 1% at most, a value that is largely surpassed in this simulation.

Existing research is generally content with providing a list of cell towers to which a device could have connected to at a given location as a result of Idle Mode surveys. Consequently, little to no information is available on how the probabilistic behaviour of mobile devices can be modelled. With the current state of knowledge, only black box simulations of the alleged behaviour in comparable conditions allow to obtain relevant data. For both versions of the fact, with sufficient repetition, the fraction of all simulations on which the same cell sites were observed as in the evidentiary data will approximate the probability of the evidence given a particular proposition.

Similarly, little research exists on what comprises «comparable conditions». Since network providers can turn off and modulate the intensity of an antenna signal based on the number of devices attempting to connect at the same time, it is widely accepted that a moment where a similar amount of people are in the area is necessary. This necessary condition is generally approximated by taking measurements during a same time frame on either a workday or weekend day (Bell, 2015).[3]

The work presented in this chapter expands upon existing research in as-

[2]Studies have shown that results from a single survey are likely incomplete and multiple measurement series, ideally with multiple devices should be made (Tart et al., 2012, 2021; Lopez, 2021).

[3]It has been theorised that comparable weather conditions (same season, precipitation, and air humidity) are required to obtain similar results because water, either as vapour, fog or rain in the air or in the leaves of trees and other vegetation, has an impact on the propagation of electromagnetic signals, (Hoy, 2015). However, to the knowledge of the author, this has never been tested empirically.

sessing uncertainties in CDR Data Analysis, in that it provides an approach allowing the evaluation of said uncertainties in light of two concurring propositions.

4.3 Discussion of the Framework

This section discusses the framework in this scenario. Figure 4.3 shows the Bayes Net for this particular scenario.

In this situation, the Bayes Net does not provide any particular added value. Indeed, the formula for the LR (cf. Formula 4.1) in this particular situation is just the probabilities of the observed evidence E (the connection to a specific cell tower) given each of the propositions.

$$LR = \frac{Pr(E \mid P_1)}{Pr(E \mid P_2)} \tag{4.1}$$

For both probabilities in this formula, a value needs to be assigned. This value should be indicative of the likelihood that the device of the accused connected to the cell tower observed in the evidence at the moment of the phone call, at the crime scene or at his home respectively. As discussed in the previous Section 4.2, with the current knowledge about the workings of these systems, this value can be approximated only through black box studies in conditions considered equivalent.

4.4 Simulation of Data

For the scenario presented in this chapter, there is no need to generate evidentiary data, as the connection to a given cell tower can be assumed among the observed cell towers. Nevertheless, some thought is put into what data is

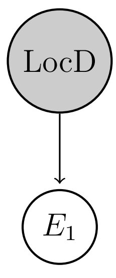


Figure 4.3: Bayesian Network for Scenario 1

assumed to be recovered in Subsection 4.4.1. In the remainder of this section, the process of choosing the locations representing the location of both propositions is described (Subsection 4.4.2) as well as the generation of reference data (Subsection 4.4.3).

4.4.1 Evidentiary Data

For this scenario, data is not simulated through an experiment. Instead, it is assumed that the cell tower visible at both locations was observed as the tower through which the call was routed at the moment of the crime. Nevertheless, some reflections must be made about the circumstances of the creation of the trace. This conceptualization is completed to help an expert address common considerations they would encounter in a real world scenario.

Considerations about the time and weather: It is presumed that the call took place on a workday during the day. The weather is assumed to be sunny without clouds, the same as the weather on the day of recording of the reference data. It is also assumed that the investigated event took place shortly before the measurement of the reference data. As such, it is unlikely that major changes in the cell network were made by the operator. It could even be assumed that cell site measurements were conducted by the crime scene investigator, confirming that no major changes in the network took place (at least in the surroundings of the crime scene).

Considerations about the device: The device is supposed to be a Samsung Galaxy S 7 (SM-G930A) running under Android version 8.0.0. The SIM card of the evidentiary device is assumed to be from the Swiss network provider «Swisscom» and running the same phone and data plan as is used for the reference data creation. Whilst the provider being the same is essential (as providers have their own distinct network), further research is required to understand whether data plans and the make and model of the device have an influence on the connections or not. It is also assumed that networking settings of the device were investigated and found to be the standard Android 8.0.0 settings.

Considerations about the evidence: It is assumed that CDR for the phone of the accused were obtained through Swisscom, the network provider with whom the SIM used in the phone was registered. In these records, a call made at the moment in time is observed routed through the cell tower

with the cell ID 6674430 as indicated in Table 4.3. This ID was chosen based on the choice of the locations described in the next section (Section 4.4.2).

4.4.2 Choice of the Locations

Two locations were chosen to allow a demonstration of LR evaluation on cell tower evidence. The two locations needed to be as such that both locations were covered by the same antenna. To this effect, measurements were conducted using a TSME 6, a device allowing to measure cell tower signals. For ease of access, signal strength was mapped on the campus of the University of Lausanne. In anticipation that SIM-cards registered to the Swiss network provider «Swisscom» were going to be used, the data was filtered for only cell towers by this provider. Manually, locations were identified where at least two cell towers were visible. Among those, two locations were chosen that are both covered by the antenna with ID 6674430 and are protected from precipitations, as the measurements were ongoing for the entirety of two days. This last criteria turned out to be a good decision. Both days of measurement were particularly sunny and being in the shadows stopped all involved devices from overheating. Figure 4.4 shows the selected locations as well as the estimated positions of relevant cell towers for this scenario.

4.4.3 Reference Data

During a two day period, single spot measurements in connected mode are conducted by simulating the claimed behaviour on devices of the same make and model as the evidenciary device. These simulations are aimed at reproducing the situation in which the trace was generated. It is unclear whether turning the devices off between each measurement had an influence or not, however, conducting the call most certainly had. Indeed, it was observed that the devices mostly connected on a 4G antenna upon startup and then switched to a 3G antenna when passing the call. It is therefore recommended to actually reproduce the activity that lead to the evidentiary data being created in real world cases as well.

Four devices of the model Samsung Galaxy S 7 (SM-G930A) from the UNIL School of Criminal Justice device park were used to conduct the simulations. All devices were running under Android 8.0.0. To differentiate the devices, the identifier from the material management system is used and the devices are called ESC-014, ESC-015, ESC-017 and ESC-018 respectively. Two SIM cards were used running a «Swisscom» data and call plan. Throughout the simulation, the same SIM-card was always used as the caller SIM and the receiver SIM. Each device had the app «Network Cell Info Lite»

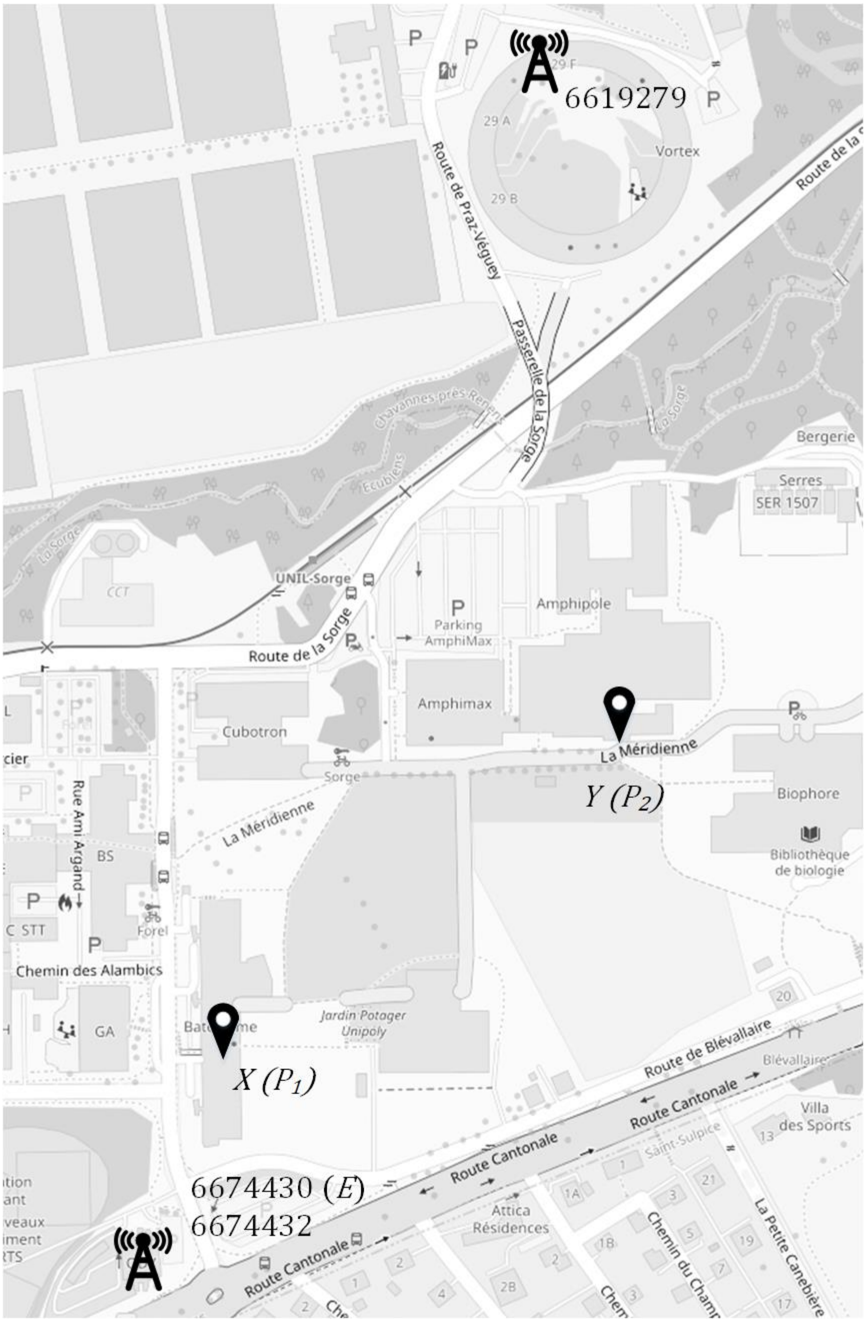


Figure 4.4: Location of measurements and relevant Cell Towers. Cell Tower position is estimated from publicly available data (swisstopo) and personal measurements. Map: OpenStreetMap

version 5.0.23 installed, an application giving information about the cellular network currently connected, allowing to obtain the information about the connected cell directly from the device itself.

For each location, two half days, an afternoon and a morning were measured. The location was changed at lunch to reduce the risk of the particular weekday having an impact. Each half day consisted of the same measurement program cycling through 4 devices. This was done to allow insight on the impact of the choice of the device as well as to ensure the battery allowing the simulation. Indeed, it was observed that the process is quite draining on the battery of the devices. Table 4.1 shows the program at each site.

Caller	Receiver	Number of calls
ESC-014	ESC-015	15
ESC-017	ESC-015	10
ESC-018	ESC-015	10
ESC-015	ESC-018	10
ESC-014	ESC-018	15

Table 4.1: Program of simulations per half day at each site. Per site, this program is conducted twice, once on a morning, once on an afternoon.

In retrospective, this plan may not have been the best choice, as temporal impacts are consequently correlated with a particular device and not the same number of calls are conducted per device. Table 4.2 show the number of calls made per device at each site.

Device	ESC-014	ESC-015	ESC-017	ESC-018
Caller	60	20	20	20
Receiver	0	70	0	50
Total	60	90	20	70

Table 4.2: Number of calls per device and location

Each simulation is conducted as follows:

- Both the receiving and the calling phone are started up
- The «Network Cell Info Lite»-app is launched on both devices
- A call is launched from the calling to the receiving phone.
- The call is accepted on the receiving phone.

- The cell tower to which each phone is connected is noted in an Excel sheet.
- The call is ended.
- Both phones are turned off.

With the devices used in this simulations, this process took about 2'30" to complete, with a slightly longer break at the end of each run as the SIM cards needed to be changed.

In parallel to the simulations, measurements were conducted with both a TSME 6 and a Snyper Graphyte LTE V3 to detect potential changes in the environment as well as to create data which might give insight into whether it is possible to generate data purely on the measurement of signal strength. As these measurements were conducted on one site per time only, a control measurement with a second Snyper LTE V3 was conducted at the other site to ensure that no particular difference is observed on the other site.

4.5 Conducting the Analysis

Using Excels «COUNTIF»-function, occurrences of each cell tower being observed were counted and frequencies calculated. The fraction is assigned as the probability value of this observation taking place at the given location.

Analysis was conducted under the following lights:

Per device: Conducting the analysis described above generates data and leaves the device open to wiping attacks as it has to be connected to the network. It is therefore preferable to use devices of the same make and model instead of using the evidentiary device. This analysis was conducted to see whether there is a substantial difference between the devices. In addition, as due to experimental design there is a substantial difference on how many measurements per device were conducted, this separation gives some insight into the robustness of the method.

Caller vs. Receiver: To conduct such experiments, two devices are always required: a caller and a receiver. If both devices give in comparable results, this will effectively allow to double the measurements per time period by simply noting the connecting cell tower on both devices. As only devices ESC-015 and ESC-018 were used as both caller and receiver, this analysis is conducted only on these two devices.

Temporal analysis: As was quite quickly evident from the measurements, there is an important temporal factor at hand with the frequency of connections varying heavily over time. To assess the impact of these variations, a temporal analysis was conducted. Additionally, this analysis gives an insight into how many measurements are necessary to obtain stable results.

Per Site: The value of primary interest for this analysis is the probability to obtain the observed result at a given location. This value is approximated by conducting an analysis per site.

4.6 Results

For almost all devices and the locations, there is not just one antenna that is systematically chosen. In the collected data, the 6674430 antenna was observed at both measurement sites as expected and, at each site there was one other distinct antenna emitting, each only being visible at their location. Figure 4.5 visualises the observed connections and Table 4.3 shows the characteristics about each cell tower. Measurements at the site revealed that all these antennas are emitting on band 8 corresponding to the E-GSM-900 technology. It therefore looks as if this particular device only chooses within a specific technology if available to pass calls. At both sites, an additional antenna by this provider emitting within this frequency has been observed in measurements, without any device ever connecting to it. The Snyper measuring device seems not particularly well adapted to detect the secondary antenna. At location P_1 it is only detected once, the one as P_2 has not been observed at all.

MCC	228 (CH)	228 (CH)	228 (CH)
MNC	1 (Swisscom)	1 (Swisscom)	1 (Swisscom)
LAC	101	101	101
Cell ID	6674430	6674432	6619279
Band	8 (E-GSM-900)	8 (E-GSM-900)	8 (E-GSM-900)
Visible at P_1	Yes	Yes	No
Visible at P_2	Yes	No	Yes

Table 4.3: Information about the cell towers observed at the locations of interest. The cell tower with ID 6674430 is the one assumed to be considered as evidence.

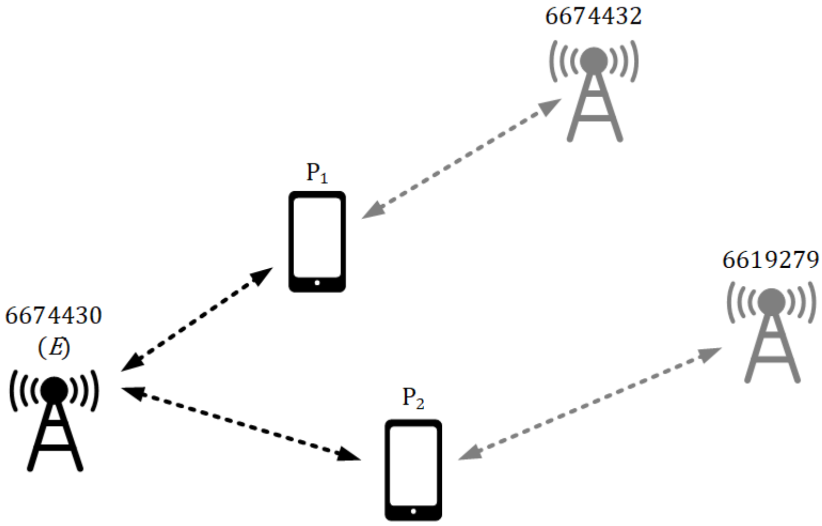


Figure 4.5: Schematic visualisation of the observed connections at sites $X(P_1)$ and $Y(P_2)$. At both sites, the evidentiary cell tower (E) was observed, as well as another one, specific to the site.

4.6.1 Influence of Caller versus Receiver

Only the devices ESC-015 and ESC-018 were used both as caller and receiver. Comparing the fractions of connection to a particular cell tower, there is no evidence that the role the device has in the call has a major influence on the choice of cell tower. Differences range from 0.01 for device ESC-015 at location P_2 up to 0.12 for device ESC-018 at location P_1. The values are shown in Table 4.4.

Device	Location	Caller	Receiver	Difference
ESC-015	P_1	1.00	0.90	0.10
	P_2	0.75	0.74	0.01
ESC-018	P_1	0.85	0.94	0.09
	P_2	0.80	0.93	0.13

Table 4.4: Fraction of devices connecting to the tower with cell ID 6674430.

4.6.2 Influence of the Device

To observe whether the choice of the device has an influence on the choice of the antenna, the obtained fractions per site is compared between the different devices. This comparison has limited value, as due to the way the simulations

were conducted, there is an intrinsic link between the devices and the time of the day. As is seen in Subsection 4.6.3, a substantial variation in time was observed, which could impact the result of the analysis at hand. The values per device are illustrated in Figure 4.6. Taking a look at the fractions per device, some difference can be observed. Table 4.5 shows the mean, standard deviation and range between the fractions. Especially the range is quite important with values differing as much as 0.21 between ESC-015 and ESC-017 for site P_2. For this present work, it is assumed that this difference is acceptable. However, further research should definitively be conducted if the here presented approach is to be used in a real world case.

Location	Mean	Std. Dev.	Range
P_1	0.90	0.06	0.13
P_2	0.84	0.09	0.21

Table 4.5: Mean, standard deviation and range of fractions per device for both locations.

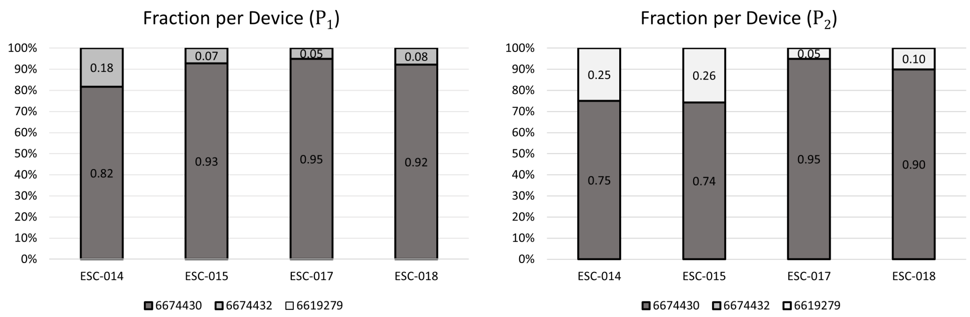


Figure 4.6: Fraction of connections per phone that connected to a particular tower at location P_1 (left) and location P_2 (right).

4.6.3 Temporal Influence

Already during the simulations, it became quickly evident that the fractions of connections to a given cell tower did not remain constant throughout the simulation period. In an attempt to visualise this temporal variability, a rolling sample over 10 simulations was chosen. For each 10 simulations, corresponding to 20 data points, the fraction of devices connecting to a given cell tower was calculated and plotted. The result can be seen in Figure 4.7. As can be observed, the fraction varies from less than 50% on the

primary tower to always connecting on the primary tower. Additionally, some similarity between the two curves can be observed. Both have a major peak for the secondary antenna around the middle and right at the end, corresponding to right before lunchtime and at around 16h00. Also, a period where the primary cell tower is responsible for all the connections can be observed for the period corresponding to the beginning of the afternoon. It is possible that these variations are due to a change in signal strength emitted by the primary cell tower, a behaviour called «cell breathing». In this case, it would be suspected that the secondary cell on each site was emitting with a signal strength closer to the primary cell during the periods where the secondary antenna peaks and significantly weaker during the period where the primary is dominant. Independent of the reason, these results underline the necessity to have data that was taken during a time comparable to the time of interest. Further research is needed to assess what «comparable» means in this context.

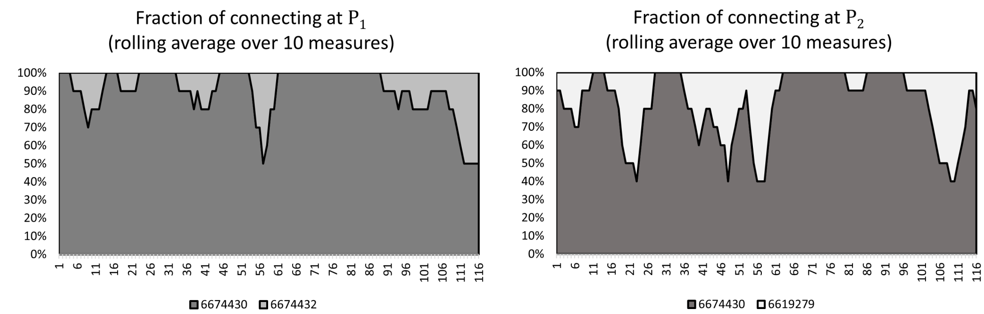


Figure 4.7: Rolling average of fraction of CT connections over 10 measurements.

A graph plotting the cumulative fraction, involving all measurements up to the current measure was created (cf. Figure 4.8). This is the value that will then be used to obtain the LR. Despite the strong fluctuations, the fraction remains relatively stable, as the large quantity of measurements is able to balance out the fluctuations. For both sites, the graph stabilises at around 30 measurements.

4.6.4 Likelihood Ratio

To calculate an LR, the fractions observed are assigned as probabilities to observe the evidence on each site respectively. When averaging over the entirety of the measured data, $Pr(E_1|P_1) = 0.900$ and $Pr(E_1|P_2) = 0.821$. This leads to a likelihood ratio of 1,1.

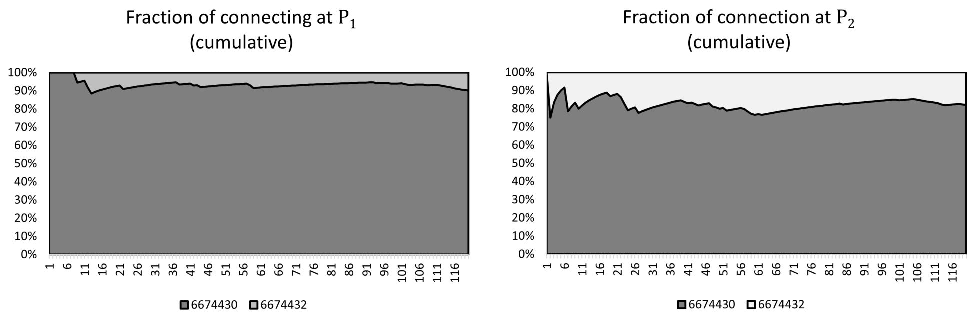


Figure 4.8: Cumulative fraction of CT connections. The value on which the graph lands on at the far right corresponding to the overall fraction is assigned to calculate the LR.

As the temporal influence is the least well controlled factor of all, the influence of time on the LR is studied. By plotting the cumulative LogLR in function of the measurements, it can be observed how quickly it stabilises at a value. The logarithm of the LR is used instead of the LR directly, as this makes the graph symmetrical around the value of non-probative evidence (0 with the LogLR instead of 1 with the LR). These two graphs are shown in Figure 4.9. On the cumulative graph it can be seen that the likelihood ratio stabilises more or less after 30 to 40 measurements. Taking a look at the rolling LogLR shows that the LR is not quite as stable as it seems from the cumulative LR, dipping in favour of the alternative hypothesis from time to time. Again, this may be explained with temporal fluctuations that seem to be more important at site P_2. However, with the LogLR being quite close to 0 anyway,the fluctuations are not massive, never even reaching a value of 1.

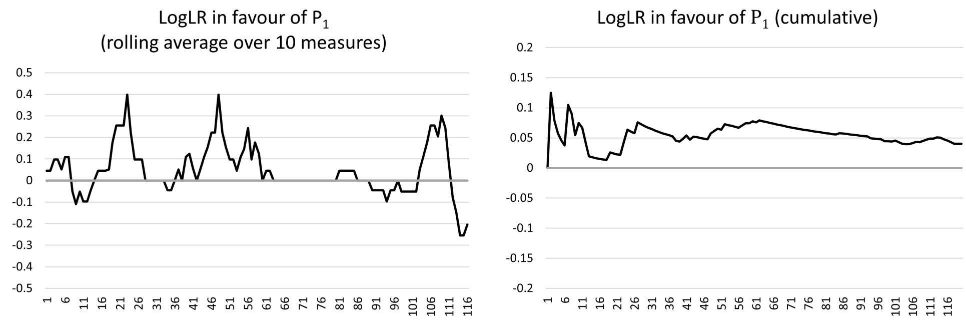


Figure 4.9: LogLR based on the fractions from the rolling average fractions (left, cf. Figure 4.7) and the cumulative fraction (right, cf. Figure 4.8).

4.7 Discussion

With the scenario presented in this chapter, a case was illustrated where a single point of cell tower evidence is evaluated under two propositions at the device level with varying locations. With the chosen locations, an LR of practically 1 was obtained, indicating that no substantial difference was observed. This is not particularly surprising, as the evidentiary cell tower was dominating at both sites. Consequentially, the behaviour of the devices used for simulation was very similar at both sites and an LR of 1 was the result that should be obtained in this situation. The results obtained do, however, indicate that it may very well be possible to obtain LR different from 1. Indeed, if as an alternative location one was chosen where the evidentiary cell tower only appeared as frequently as the secondary towers did in this simulation, an LR of around 10 would be the result. This still is not an extraordinary value. In the verbal scale proposed in (Marquis et al., 2016), this would be qualified as limited or weak support for one of the hypotheses over the other, but it still constitutes support.

The simulations conducted in this work show an abundance of factors influencing the obtained result that are not properly understood. To this day, it is quite complicated asserting that two situations can be considered «comparable» and that measurements therefore are relevant for the situation of interest. Notably, further research is required to understand temporal factors. Additionally, it would be beneficial to conduct research into the possibility of predicting connection fractions based on signal strength measured by survey devices, as this would render conducting such analyses far less time consuming.

Chapter 5

Scenario 2

The scenario presented in this chapter illustrates a question on the level of the device user. Questions related to location are not touched upon. Instead it is shown how behavioural biometrics on a device can be evaluated in light of propositions of user identity.

5.1 Description of Scenario

The aim of this scenario is to show the reasoning on the level of the user. The question of interest is whether for a given day, the habitual user of a device was utilising it, or whether another, specified person was using it on the day in question. A situation, where such a problem may arise is described below:

> Digital documents containing trade secrets were stolen from an enterprise. The theft was traced back to a particular smartphone from the enterprise's mobile device pool. The person who normally used this device claims to have lost her device on this day. She asserts that the device was given back to her by another employee, which she names as a suspect.

In such a situation, it can be envisioned that the devices of both users were seized and analysed by a forensic expert. For the sake of this scenario, it is supposed that a tribunal tasked the expert to conduct an analysis based on usage patterns to determine which of the two persons was the one using the device during the day of interest. The observations are to be evaluated in light of the following two propositions:

P_1:	Person A was using the device Y at time t.
P_2:	Person B was using the device Y at time t.

5.2 Theoretical Background of Behavioural Biometrics

Identifying perpetrators based on their behaviour is nothing new. In crime analysis, the *modus operandi*, the way an offense was committed, is used regularly to link crimes committed by a same offender. For example in burglaries, the choice of target, the time of the day as well as the way the targeted building was entered present efficient ways to identify cases where the same intruder was the perpetrator. This approach is based on the idea that a serial offender will act in the same (or at least similar) way, different to other offenders, every time he commits an offence. In crimes with a highly serial character, such as burglaries, the *modus operandi* has been shown to be sufficiently specific to identify series (Ribaux, 2014). In a first step, this approach does not associate a civil identity with the entity of the perpetrator or the perpetrators, but if at any point in time, this link is made, the entirety of the series may be resolved instead of just one case.

A similar logic is followed with the analysis in this chapter. For a day of interest, the way a smartphone is used is analysed, and then compared to other days within a reference period. If a high correspondence in the way the device is used is observed, this is an indicator that the device was used by the same person. Again, by itself, this analysis does not identify which person was using the device. However, if the identity of the user during the reference period is known and uncontested, then an opinion can be expressed on the physical person who was using the device on the day of interest.

Technical solutions that are based on behaviour as a means of identification have been proposed and used as means of continuous authentication in an IT-security context. The idea is that a particular characteristic of user behaviour is constantly observed. If the behaviour of the user at some point in time is outside the expected range of behaviour, the device locks itself and thus prohibits a different person from accessing its content. There are three approaches to performing this process (Al Solami et al., 2010):

1. either a training set for the legitimate user and a potential adversary is available

2. a data set is available only for the legitimate user

3. the reference data is constantly generated and only abrupt changes in behaviour are looked for

Examples of characteristics that have been studied include keystroke dynamics (Bhatt and Santhanam, 2013; Saevanee et al., 2015), the usage of

applications (Li et al., 2011) or the language used when writing messages (Saevanee et al., 2015).

So far, few concepts of behavioural biometric authentication for mobile devices have established themselves as widely used real world applications. Solutions exist for developers to improve authentication through behavioural biometrics (LexisNexis, 2020). On the user side, some services offer application based on swiping over the screen in a particular way as an unlock mechanism (Q Locker, 2022; Lock Screen Master, 2022), but other than that, most approaches have not been popularised yet.

In a forensic context, a method proposed in (Guido et al., 2016) is of particular interest, as, whilst developed for continuous usage, it is based on data that is available in a forensic extraction of a mobile phone. This approach was reproduced and adapted for usage in a forensic setting in the master-thesis of Michelet. He proposes two approaches in his work, one focusing on classification between two users and one where the distance of a vector of characteristics is used to quantify the difference in behaviour for two individuals (Michelet, 2021). Based on the latter of the two approaches, a slightly modified process is proposed here allowing for the generation of an LR as an outcome of the analysis.

5.3 Discussion of the Framework

The propositions at hand are identity-focused and have a person-level subject. The Bayesian Network is adapted in order to represent the reasoning of the scenario. All location-related nodes are removed. The only evidence node is derived from «UseU». The evidence observed is the evidence score between the activity observed in the period of interest and the period of comparison. Figure 5.1 shows the adapted Bayes Net. Whilst technically, with the parameters of the scenario at hand, the nodes «User» and «UseU» could be left out and the evidence could directly be connected to «UseP», they were left in to make explicit the reasoning in mounting to a person-level subject.

Additionally, given the setting of the scenario, the possible states of the nodes have been adapted, removing the possibility of «no user» and «someone else», but adding a state for Person B as the user in node $UseP$. In the node «User», the prior probabilities of Person A being the general user $Pr(User1)$ is set to 1 as general usage is not contested by either party.

In most Identity-focused scenarios in forensic science, the alternative proposition considered is one where an open set of alternate sources are considered. Generally, P_1 considers person A to be the source and P_2 considers

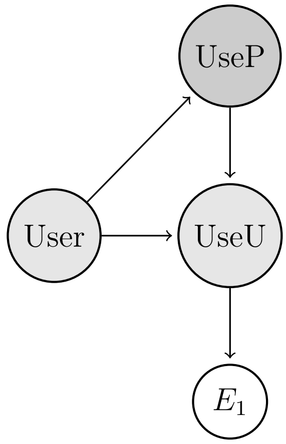


Figure 5.1: Bayesian Network for Scenario 2. Conditional probability tables remain the same as indicated in chapter 3.

the source to be not person A, but anyone from a non-nominative population of alternative persons considered to be relevant for the case at hand. In such a situation, an LR can be obtained through comparing the intra-variability of A (the probability to observe E if A is the source) to the inter-variability of A[1] (the probability to observe E if A is not the source). In a Scenario where person A is compared directly to person B however, there are four ways an LR can be obtained:

1. Comparing the intra-variability of A to the intra-variability of B.
2. Comparing the intra-variability of A to the inter-variability of A.
3. Comparing the intra-variability of B to the inter-variability of B.
4. Comparing the inter-variability of B to the inter-variability of A.

If probabilities were inherent to an event and it would be possible to know them, these approaches would be identical. Indeed, in a situation where only two entities are considered as a potential source, $Pr(E \mid \overline{A})$, the probability of observing E if A is not the source, and $Pr(E \mid B)$, the probability of observing E if B is the source, should be identical. However, in the real world, if the probability of the evidence can not be given directly, it has to be assigned based on an approach approximating the probability. This has to be done based on a limited set of observations and a process that will create different

[1]Technically, this is the intra-variability of the observed characteristic and not the intra-variability of the person. This shortcut is made to improve readability.

results depending on whether one observes the intra- or the inter-variability. As a consequence, the probabilities will differ based on the approach chosen. As this problem arises only through the imperfections of the real world, there is no hope in finding a solution through theoretical considerations about the interactions of the probabilities. An argument can be made for using the second variant, as it is the approach used when comparing to the general population and having one person as the alternative, the present situation is basically just a fringe case of the general population where the size of the population is 1. However, this poses an immediate challenge, as Person A and Person B should be interchangeable without impacting the result. Approach 2 and 3 should therefore produce the same result (which they will likely not).

The author is of the opinion, that approach 1 should be prioritised, as we intuitively expect the probability to observe an element if it comes from a given source to be independent of what alternative sources are proposed. Indeed, if the scenario were to change and instead of person B a third person C is proposed as the alternative source, in approaches 2 and 4, $Pr(E \mid P_1)$ would have to be reevaluated despite P_1 not having changed. The same issue appears when the first proposition is changed for approach 3. Approach 1 is the only one where this is not the case and is therefore followed for the present analysis.

5.4 Simulation of Data

To generate needed data, two volunteers were each given an iPhone 6s (A1688) running iOS 14.4.4. The volunteers (Person A and Person B) were given one day to prepare their device, Phone 1 and Phone 2 respectively, in order to reduce the influence of setting up the device on the data. Subsequently, the volunteers used their device for three consecutive weeks as their principal device. Then a full file system extraction was conducted using Cellebrite UFED (Version 7.53). Phone 2 was then given to Person A who, after another day of setup, was using the device regularly for a full day before an extraction was conducted on the device again. In this manner, two sub-scenarios were created: For sub-scenario S1, the last day of the three week period of Phone 1 was considered as the day of interest. In S1, the ground-truth is that Person A used his own device to download the stolen data. For sub-scenario S2, the additional day of Person A using Phone 2 is considered as the day of interest. In this scenario, Person A uses Phone 2 (the phone of Person B, which is the accused) to access the data that was stolen. In both scenarios, Person A is the person having committed the theft, in sub-scenario S1 using their own

device, in sub-scenario S2 using the device of another person, Person *B*. The calendar of the measurements is shown in Table 5.1.

Week	Weekday	Mo	Tu	We	Th	Fr	Sa	Su
1	Phone 1	S(A)	A	A	A	A	A	A
	Phone 2	S(B)	B	B	B	B	B	B
2	Phone 1	A	A	A	A	A	A	A
	Phone 2	B	B	B	B	B	B	B
3	Phone 1	A	A	A	A	A	A	A
	Phone 2	B	B	B	B	B	B	B
4	Phone 1	E_{S1}(A)	Ex					
	Phone 2	B	Ex		S(A)	E_{S2}(A)	Ex	

Table 5.1: Calendar of the data simulation period.
S(A)/S(B) = Set up Day for person A/B;
A/B = Reference Day for person A/B;
Ex = Extraction of the phone;
E_{Sn} = Day of interest for Subscenario n

5.4.1 Application in a Real World Scenario

Using the approach here presented in a real world scenario is challenging. First of all, a reference period needs to be fixed for which it is agreed that the same user is using the device. This may not always be evident or even possible. Additionally, this reference period should be for comparable circumstances, notably regarding workdays versus vacation, not encompass sick days or days with a completely different workload. This on its own may prove to be an impossible task to resolve, especially as there is little to no research into what factors may be influencing whether two periods are comparable or not.

Until influencing factors are better understood through future research, the author considers that application of the present method in a real world case should be done only with significant reservations and care.

5.5 Conducting the Analysis

This section describes how the analysis was conducted, first discussing the used characteristics, second describing the mannre used to obtain probabilities.

5.5.1 Choice of Characteristics

Based on the work conducted in (Michelet, 2021), a distance-based approach exploiting only system data was chosen for this work. This approach was chosen for multiple reasons: First, the distance-based approach gives as a result a one-dimensional measure for similarity: the distance. Based on that, a distribution under each proposition at hand can be generated with relative ease. While the classifier used in Michelet's first approach does generate probability-scores, it is unclear as to how these scores came to be, what their meaning is and how they are to be interpreted. Second, the distance-based approach can more easily be generalised on populations rather than a single person as the alternative population. This would be somewhat more difficult for the classification approach, as the model would need to be trained on the alternative population consisting of the data of multiple other persons. In addition, it is unknown as to how the model would react if the "someone else" postulated from the alternative proposition, in this case B, is not actually part of the population. This is not an issue with the distance-based approach as scores under the alternative proposition just give an answer to the question "What distance would we expect if someone else were to have used the device on this day?". This allows for the same approach to be reused in scenario 4 described in Chapter 7, where the person of interest is compared against an open set of alternative users.

The focus on system-artefacts[2] only has several advantages:

- As shown in (Michelet, 2021), system artefacts provide as good as, and in some cases even better, results than higher-level artefacts.
- System artefacts should be less dependent on the usage of specific applications than higher-level traces such as app usage.
- System artefacts do not contain personal information. As such they are far more innocuous from a data protection perspective, which provides an advantage when handling the data of experiment subjects.

Using only those characteristics, only the two following files are recovered from the conducted extraction:

- knowledgeC-database located at /private/var/mobile/Library/CoreDuet/-Knowledge/knowledgeC.db and its associated wal- and shm-files
- Lockdown-logfile located at private/var/logs/lockdownd.log

[2]The system-artefacts considered by Michelet consist of characteristics resulting from system functions of the device.

From these two files, 90 characteristics are created for each day. These characteristics contain information about power-on events, the display orientation, whether the device is plugged in, the screen illumination, the lock-state, airplane mode, WiFi and Bluetooth connections, battery state, Siri, media playing and app usage. A full table of all the characteristics can be found in Annex C.

5.5.2 Estimation of the Probabilities

The probabilities of interest in this scenario is the probability to observe the behavioural characteristics at the date of interest if the person was Person A ($Pr(E \mid P_1)$ versus if the person was Person B ($Pr(E \mid P_2)$). To estimate this probability, the same behavioural characteristics are observed for a reference period for both persons (cf. Section 5.4). Using the anonymisation-script from (Michelet, 2021), system-level behavioural characteristics are recovered from the files of the two iPhone extractions. The characteristics for each day are stored as a 90-dimensional vector. As the scores for the different variables vary quite heavily in order of magnitude, the vectors are normalised using the parameters of the entirety of the vectors in the reference data. A principal component analysis is conducted on the data and the first five principal components (PC) are studied in order to assess their suitability as a separating characteristic. Consecutively plotting the first PC against one of the other PCs revealed that PCs 2 through 5 did not add anything to the separation of the two populations at hand. This is shown in Figure 5.2 Based on this analysis, the first PC was chosen as a singular indicator. An intra-variability-distribution for both populations is created by sampling 15 days out of the reference period repeatedly. One day is randomly chosen as a "day of interest" and the distance between the PC1-value of this day and the center of gravity of the PC1-values of the remaining 14 days is calculated. To see how well the two population separate between each other, the inter-variability-distribution is generated, with the comparison element originating from the population of P_2. The same process is repeated for the data created by person B (P_2).

By transforming the vector for the day of interest in the same way, a value is obtained as the evidentiary value: the vector is normalised using the means and standard deviations obtained from the reference data, then treated using the parameters obtained from the PCA conducted for the reference data. The distance of the first PC to the mean of the values[3] obtained for the first

[3]Given that only one dimension is looked at, the mean corresponds to the center of gravity.

PC of each population are the values taken as observed values under each proposition.

Using the fitter function from the python «Fitter» module, a density distribution is fitted to the data. Evaluating the density function at the evidentiary value observed, the probability of the evidence given the population at hand can be obtained.

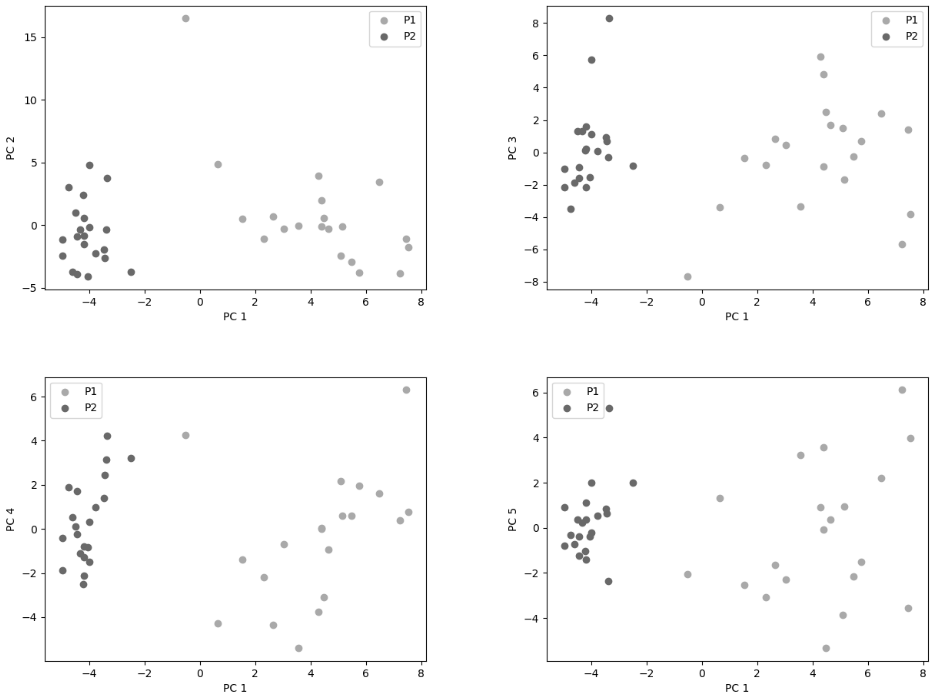


Figure 5.2: Plots of the first vs the second (top left), third (top right), forth (bottom left) and fifth (bottom right) PC. As can be seen, P1 and P2 separate out well based on PC1 without the other PC adding anything further to the separation.

5.6 Results

Plotting the histogram of the intra-variability and the inter-variability for the distances of both populations shows a good separation between the two populations, as would be expected based on the PCA plots. This is shown in Figure 5.3. Fitting a distribution over the intra-variability values, an exponential was obtained for P_1 and a beta-distribution was obtained for P_2.

These distributions were chosen as they resulted in the lowest sum of square errors and are shown in Figure 5.4.

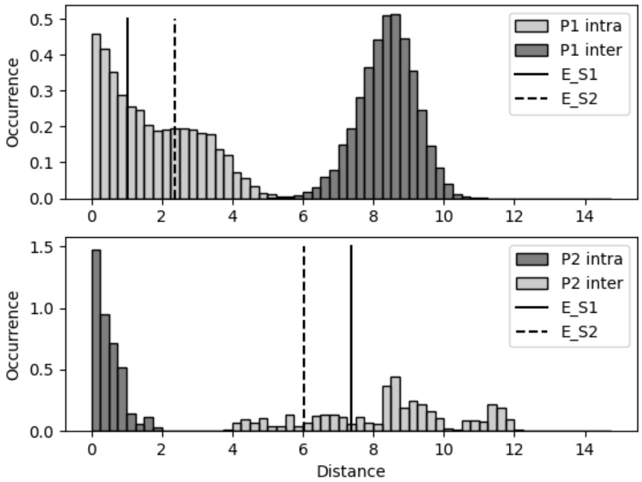


Figure 5.3: Histogram of distances observed for intra- and inter-variability of P_1 (top) and intra- and inter-variability of P_2 (bottom). The values obtained for both E are indicated as lines.

	E (S1)	E (S2)
$Pr(E \mid P_1)$	0.327	0.193
$Pr(E \mid P_2)$	0.001* (3.414×10^{-07})	0.001* (6.653×10^{-06})

Table 5.2: Probabilities assigned for the evidence given the propositions for both scenarios.
*: values were lower bound at 10^3. Density values obtained from the distribution are indicated in brackets.

The values obtained under the alternative hypothesis are rather small. They are obtained from a region of the distribution where no actual values were observed. Whilst it is justifiable to use those values directly, there is also a strong argument to be made, that these values are far too impactful given the low quantity of data available. Indeed, at the far ends of a given density function, values may easily vary several orders of magnitude depending on very little. To address this issue, the expert may assign lower bound values under which the values of the distribution are ignored and the lower bounds value is taken as a probability instead. In the current situation, the author is of the opinion, that values below 10^{-3} are not justified given there were only

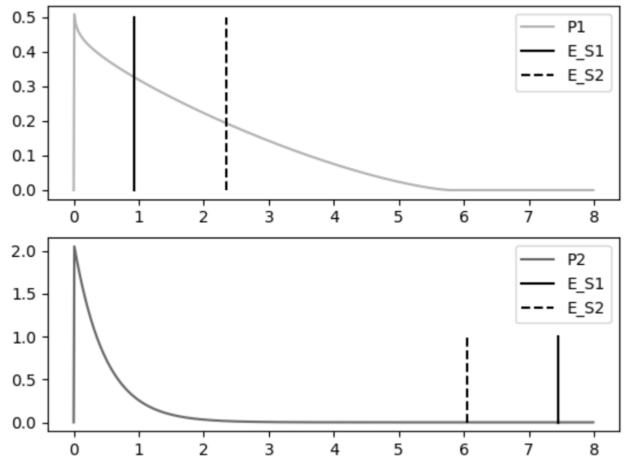


Figure 5.4: Density distributions for the intra-variability of P_1 (top) and P_2 (bottom). The distances observed as evidence are indicated as vertical lines

21 initial observations for the alternative proposition, although the sampling has allowed to simulate a larger population. Lower bounding the probabilities ensures that no astronomically high LR are presented that have no sufficient support in data. A disadvantage of this approach is that once a certain level of dissimilarity reached, the similarity of the evidence with the reference element does not really have an impact on the LR anymore. This can be seen by looking at the two subscenarios presented here. Despite E_{S2} being more similar to P_2 than E_{S1}, the probability under P_2 that is assigned ends up being the same in both cases.

Inputting the values into the formula for the LR or the Bayesian Network, the following LRs are obtained:

	S1	S2
LR	327	193

Table 5.3: LR for both subscenario.

Following the verbal scale by (Marquis et al., 2016), both these LR are qualified as strong support for the proposition that person A was the person using the device at the day of interest.

5.7 Discussion

LRs of 327 respectively of 193 in favour of person A using the phone during the period of interest are obtained in this chapter. These LRs support the correct proposition, person A indeed having been the user of the device at both days of interest. The analysis conducted technically only compares the usage of the device at the day of interest to the usage during the period of reference, which would result in propositions of the form "At the day of interest, the person usually using phone 1 (respectively 2) has used the device of interest." However, given the circumstances of the case, it is not contested who the person was that used each phone in the reference period (person A and person B respectively), allowing the expert to express an opinion on the person-level and not just on the user-level. This reasoning is categorical, requiring no further probabilistic evaluation. The LR on the user-level is therefore identical to the LR on the person-level.

The characteristics used in this chapter were reused from (Michelet, 2021). It is likely that quite a degree of co-dependency exists between some of the variables whilst others do basically never change. This would mean that it is possible to reduce the number of analysed characteristics rendering the process more efficient. However, a larger data set with high diversity of participants would be required to conduct the necessary analysis.

Whilst the here presented method may provide interesting insight if a high degree of similarity is observed between the reference period and the period of interest, there is limited use for the method if differing results are obtained. Indeed, there may be reasons why a person changes their behaviour, such as them going on vacation, changing job or getting heavily invested in a new app. Research from psychiatric research suggests that based on similar characteristics, it may be possible to predict phases of mental illness (Ben-Zeev et al., 2015). It is to be expected, that changing behaviour due to mental illness would also impact the analysis presented in this chapter. Finally, as most offenders are single time offenders (Kuhn, 2012), the moment in which they do commit a crime is per definition a time frame of unusual behaviour. Future research should focus on addressing these questions.

No calibration of the used density distribution was conducted. In further research, this should be considered.

Chapter 6

Scenario 3

In this chapter, a scenario with a location-focused question is presented, which involves direct evidence of a user being in possession of the device. Two elements of evidence, one for location and one for user identity, are evaluated together under propositions of the location of a person.

6.1 Description of Scenario

In this scenario, it is shown how direct evidence of device possession allows to evaluate the evidence under person-level evaluation. Here, the parties do not just disagree on the location of the device but also on who had the device at the moment in time. The following scenario may describe such a situation:

> A crime was committed at Location X and time t. A suspect, Person A, is arrested some time after and his smartphone is seized for extraction. On the device, a picture is found showing a finger with visible friction ridge patterns, localised at the crime scene and timestamped around the moment of the crime. The suspect insists that their phone was stolen during the period in which the crime was committed and that he only found the phone the next day by pure coincidence. Person A claims that during the period of which the crime was committed, he was at home sleeping (Location Y).

The claim by the suspect may seem outlandish. This should however not impact the analysis conducted by the expert in this case, as the plausibility of an advanced proposition is in the domain of the court and is represented in the prior probabilities of said proposition. A structured Bayesian approach will allow to counter biases one may have against believing in such a proposition.

Other circumstances in which the present approach can be employed may arise when no specific information about device possession is available (e.g. when the suspect refuses to make a claim regarding possession during the period of time, or when the person that would have had this information is deceased), or when one of the parties wants to preempt an eventual claim that someone else was in possession of the device at the moment of the crime.

Either way, the following pair of propositions is considered:

P_1:	Person A was at Location X at time t.
P_2:	Person A was at Location Y at time t.

6.2 Theoretical Background of Location Traces

The global positioning system (GPS) is a network of satellites equipped with a very precise atomic clock and knowledge of their own position, perpetually verified through a network of ground-stations. Transmitting their position and their current time ($[x_i; y_i; z_i; s_i]$ for Satellite i), they allow devices that have vision of at least four GPS-satellites to calculate their own position in 4 dimensional timespace.

To understand these calculations, a simplified model can be considered, where the speed of light in the atmosphere is considered to be equal to the speed of light in a vacuum (c) and relativistic effects are ignored. In this model, the distance between a satellite and the device can be expressed through two ways. First, based on the time travelled, where t_i is the time of reception of the signal by the device and b is the bias of the devices clock in comparison to the GPS-time :

$$d_i = (t_i - b - s_i)c \tag{6.1}$$

Second, based on the geometrical distance to the satellites using Pythagoras' theorem where x, y and z are the spatial coordinates of the device:

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} \tag{6.2}$$

Obtaining a package $[x_i; y_i; z_i; s_i]$ from at least four distinct satellites allows a device to solve for x,y,z and b and therefore localise the device. If more than four satellites are available, the system becomes overdefined, leading to the need of approaches to mitigate differences[1]. Given that measurements

[1] For four satellites, the equations give a single possible solution, which is likely less correct

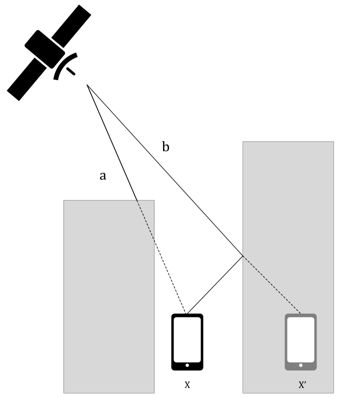


Figure 6.1: Visualisation of GPS masking and multipath: Because the direct signal from the satellite is blocked (path a), the phone receives the signal reflected from a nearby building (path b). As the calculations assume direct line of sight, the phone will localise at position X' instead of X.

are subject to error, having more than four satellites can help increase the precision of the system (Blewitt, 1997).

A major source of systematic errors with GPS are effects of multipath and masking, prevalent mostly in large cities. If a device attempting localisation through GPS is standing in the space between two high objects, such as buildings or mountains, it can happen that one of those objects blocks the direct line of sight to a GPS-satellite. If the surface of the second object is sufficiently flat, the signal sent out by this satellite may bounce off of this surface and still reach the device. This reflected signal has taken a longer path than a direct transmission would have taken and therefore a longer time of flight. As it is assumed that the signal has travelled in a straight line, the device appears to be further away from the hidden satellite than it actually is, causing the computed localisation to be at the wrong position (Van Sickle, 2020). Figure 6.1 schematically shows the mechanism behind the shadowing and echo-effect.

Nowadays, other technologies are used to improve upon localisation by GPS. Notably, so called fingerprinting techniques are employed. These approaches are based on huge databases containing lists of cell towers and WiFi access points that are visible at a given location. Reverse-lookups of these databases allow a device to improve the accuracy of its position even if only a small number of GPS-satellites are visible. Especially in cities, where a large quantity of WiFi networks are available, the precision can be improved substantially through fingerprinting (Cedergren, 2005). The technology is, however, dependent on the accuracy of the reference databases. These databases

are kept updated through the measurements of devices using them, which has shown to be challenging to correct errors once they are in the database, sometimes causing errors exceeding what would be expected from GPS-only locations.

This effect has been documented in studies looking at real world systems. Over the period of one year, Merry and Bettinger conducted a large number of measurements using the same phone at six precise locations surrounding their faculties building. They were the first ever to record a directional bias in one of the measurements they conducted. This particular location was situated on the side of their building towards the parking lot. They theorised that the observed effect may have something to do with echo from the building (Merry and Bettinger, 2019). Such effects were also reported in Ryser and Jacquet-Chiffelle (2021), where the accuracy of the geolocation associated to images was investigated. In their work, Ryser and Jacquet-Chiffelle not only observed errors up to 27km for some locations, they also reported that these errors are heavily biased in specific directions. They conclude, that depending on the location, error of localisation may vary heavily (Ryser and Jacquet-Chiffelle, 2021).

The source of these systematical errors is not definitely known, although the theory has been proposed that it is caused by a bias towards locations where many people are, caused by databases being more frequently updated at these locations than where fewer people are.

6.3 Theoretical Background of Finger Marks

Increasing the capacity to grip, friction ridge skin on the hands and feet is a feature shared by all primates (Berry and Stoney, 2001). The way they are generated, a chaotic process involving stretching and pulling during the development of the fetus, causes friction ridge patterns on hands to have a very high degree of variation between not only different individuals but also between different fingers of a same person. Additionally, due to the way the dermis regenerates, the pattern is generally stable, set aside major injuries or illnesses afflicting the skin (Champod et al., 2017). This combination makes finger friction ridge pattern a characteristic well suited to identify an individual. Whilst earlier anecdotal evidence suggesting awareness of the individuality exists, the scientific foundations for the use of fingerprints as a means of identification came about in the middle of the 19th century, when multiple British researchers more or less at the same time became interested in the properties and use of fingermarks both for the identification of persons, mostly criminals, and as a means of solving crimes through traces

left on crime scenes (Berry and Stoney, 2001). As the very act of touching is likely to leave behind traces of high discriminatory power, dactyloscopy, the discipline of identifying a person based on their finger friction ridge pattern, has become a poster domain of forensic science.

A dactyloscopical analysis is done following an ACE-V[2] methodology. This process begins with the trace and a reference print produced from a suspected source's finger being analysed on three levels:

- The general pattern, which is the macro-structure of the ridge flow.
- The *minutiae*, which are points where ridges end or merge or other macrofeatures.
- The positions of pores and other micro-features within the ridges.

These characteristics are then compared between the trace and the reference print. Both correspondences and differences are indicated, whereas an inexplicable difference systematically leads to an exclusion of the print's owner as the source of the trace. The entirety of agreements and differences are evaluated under a pair of concurring propositions, generally in the form outlined below, before finally being verified by another expert (Champod et al., 2017).

P_1:	Person A is at the source of the trace.
P_2:	Someone else is at the source of the trace.

Automatic systems for the evaluation of fingermark analysis exist, according to Champod et al. (2017) generally following one of two approaches. The first consists in creating a statistical model allowing to quantify the rarity of a given configuration of *minutiae*, as for example proposed by Neumann et al. (2011) at the Forensic Science Service for the UK and Wales. A second approach is based on comparing the analysis to the results of comparisons with known sources. Such an approach is for example followed in Egli (2009) which is at the basis of the evaluative functionalities of the tool «PiAnoS», the solution created and hosted by the School of Criminal Justice at the University of Lausanne (Furrer et al., 2020) and used in this work. The approach is based on using AFIS[3]-scores as a measure for similarity, upon which probability distributions are based (Egli, 2009).

From a security perspective, finger print scanners for unlocking digital devices have become a frequent feature of smartphones and computers. Unlike classical finger mark analysis, finger print scanners to unlock devices are

[2] Analysis, Comparison, Evaluation, Verification

[3] Automated Fingerprint Identification System

generally not based on a picture of the ridge pattern, but are based on a scan of the 3-dimensional features of the ridge pattern. The scan is then compared to reference scans stored within the device, and, if a sufficient level of proximity is achieved, the device is unlocked.

In recent years, identifying someone through their friction ridge pattern has extended further into the digital dimension. With the evolution of digital cameras built in smartphones reaching a resolution sufficient to distinguish the ridges on the hands visible in pictures taken and cases have occasionally happened where such evidence was considered. Two early cases where such an analysis was done are a 2015 identification conducted by the FBI to identify a perpetrator of child abuse in Georgia (USA) from fingers visible in the illegal material (FBI, 2018) or a drug case investigated by the South Wales Police (UK) in 2018, where finger mark experts were able to identify a drug dealer based on a picture of him holding the wares offered for sale (Wood, 2018).

6.4 Discussion of the Framework

When direct evidence of possession is available, there will in most cases not be much use to discuss general ownership of the device. Evidence of a given physical person having the device at a given moment in time will generally be more impactful than indirect evidence. Whilst there is no direct screen-off effect, as there is no categorical conclusion of ownership, the impact the evidence has on the overall LR will be overshadowed by the direct evidence. Consequently, the nodes on overall usage ($User$) and abstract usage ($UseU$) are removed from the Bayes Net used in this chapter. Based on the analysis conducted in Section 3.3, $Pr(UseP_1)$ is fixed at 0.8. Figure 6.2 shows the reduced Bayes Net.

Given the directional bias observed in real world systems (Merry and Bettinger, 2019; Ryser and Jacquet-Chiffelle, 2021), it is proposed that GPS evidence is best evaluated using a two step approach: First, the direction in which the measurement is situated from the proposed location is considered. Second, the distance of the measurement to the location *given this particular direction* is taken into consideration. This leads to the following LR where φ is the angle of the direction from the north and d is the distance from the location considered in propositions 1 and 2 respectively (cf. Figure 6.3 for an illustration of the two parameters):

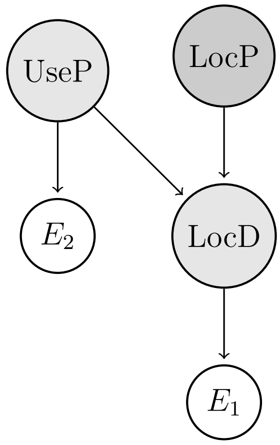


Figure 6.2: Bayesian Network for Scenario 3

$$
\begin{aligned}
LR_{E_1} &= \frac{Pr(E_1 \mid LocD_1)}{Pr(E_1 \mid LocD_2)} \\
&= \frac{Pr(d_1; \varphi_1 \mid LocD_1)}{Pr(d_2; \varphi_2 \mid LocD_2)} \\
&= \frac{Pr(d_1 \mid \varphi_1; LocD_1) Pr(\varphi_1 \mid LocD_1)}{Pr(d_2 \mid \varphi_2; LocD_2) Pr(\varphi_2 \mid LocD_2)}
\end{aligned}
\tag{6.3}
$$

As such, d and φ can be considered as two separate, although not independent, traces and the Bayes Net can be adapted consequently as shown in Figure 6.4. As the distance and angle are dependent on the coordinates of the location considered in each position, the values φ and d are different in the numerator and the denominator. This may seem counter-intuitive. It can however be shown that the above term is equivalent to the original LR formula.

As it is easier to condition a continuous distribution by a discrete measure, the angle φ is not evaluated as the probability of the angle being exactly φ for a given position. Instead, this probability is approximated with the probability that the observed angle for a localisation created at a given position falls in a section of the circle $[\varphi - \varepsilon; \varphi + \varepsilon]$. A distribution of d is then created for all simulated data points that fall within this circle segment. An ε of 30° is used in this work, giving an overall wedge size of 60°. The wedge size is chosen intentionally quite large to ensure sufficient data for the modelisation of the distance is available. Smaller angles are just as justifiable, for larger angles, care must be taken for the analysis to still be meaningful. This approach is visualised in Figure 6.5.

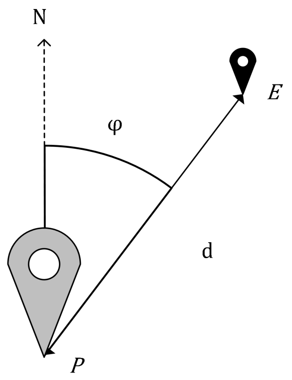


Figure 6.3: Illustration of the parameters d and φ for a given position (P) and observed localisation (E). d is the distance between P and E and φ is the angle between the $\overline{PE}$-vector and north.

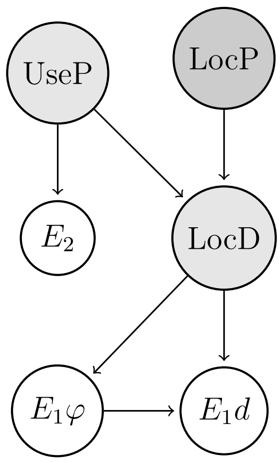


Figure 6.4: Bayesian Network for Scenario 3 adapted to take into account both elements of E_1 separately

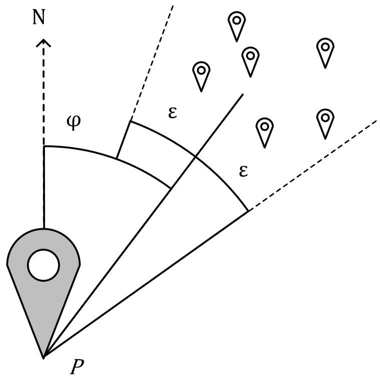


Figure 6.5: Illustration of data points retained for a given location P. Instead of evaluating φ, the probability of data points laying within a sector of $[\varphi - \varepsilon; \varphi + \varepsilon]$ is evaluated.

6.5 Simulation of Data

This section describes the creation of all the data required to conduct the analysis at hand. Given this is a simulated case, many parameters could freely be chosen. What considerations should be made when conducting such an analysis in a real world case is indicated in Subsection 6.5.4.

6.5.1 Choice of the Concurring Locations

Two locations were chosen on the campus of the University of Lausanne. The first location, considered under P_1 is within the localities of the School of Criminal Justice and represents Location X. The second location considered under P_2 is in the cafeteria area of the neighbouring building and represents Location Y. There is a distance of about 200m between the two locations.

The coordinates of locations X and Y are measured from the interactive map of the University of Lausanne campus (University of Lausanne) and are indicated in Table 6.1.

	Longitude	Latitude
P_1: Location X	6.575116326	46.521954786
P_2: Location Y	6.573832039	46.521592273

Table 6.1: Coordinates of the positions considered in each proposition.

6.5.2 Simulation of the Evidence

A picture showing the left index finger of the author was taken with a Samsung Galaxy S20 5G whilst location services, mobile data and WiFi were turned on. The picture was taken at Location X, within the localities of the School of Criminal Justice at the University of Lausanne. When the picture was taken, an active Swisscom SIM card was in the device and it was connected to the WiFi network.

6.5.3 Simulation of Reference Location-Data

Reference data was generated at the locations described by both propositions. The same phone used to create the evidentiary data was used with the exact same settings. The phone was set on a table and approximately every minute, a picture was taken. This was done at both sites over a period of 3 days, with the first two days being Wednesday and Thursday of the same week, and the third being Monday of the following week. Overall, 1627 pictures were taken. A logical extraction of the device was conducted with Cellebrite UFED 7.53.0.24 extracting only pictures. The result was opened in Cellebrite Physical Analyser 7.54.1.7 and an Excel-report was generated from the locations tab. Manually, this report was split up into two Excel lists, each one containing the reference data points for one of the locations indicated in the propositions.

6.5.4 Considerations for a Real World Case

In Chapter 4 it was mentioned that little is known about the factors influencing the choice of cell tower, even less is known about the factor influencing localisation precision. To preserve evidence, it is not recommended to use the evidentiary device in a real world case. Given that the systems providing these locations are mostly proprietary, owned and operated by Google in the case of Android- and by Apple in the case of iOS-Systems. As such, it is imperative to use a device from the same ecosystem as the evidentiary device when simulating the data. Given that the sensitivity of a device in regards to WiFi- and CT-signals is likely to vary from one model to another, it is also recommended to use a device of the same make and model. Also, whether or not these signals are detected or not is dependent on the devices connectivity settings. As far as possible, the settings of the device at the moment of interest should be recreated. In Merry and Bettinger (2019), no obvious temporal influence was observed, nevertheless, further research should be conducted to strengthen these results.

6.6 Conducting the Analysis

This section describes the analysis process for both the localisation evidence and the fingermark evidence.

6.6.1 Location Evidence

The coordinates recorded in the EXIF-data of the picture taken at Location X are shown in Table 6.2.

	Longitude	Latitude
$E1$	6.5750922	46.5219326

Table 6.2: Coordinates recovered from the Evidence E_1

The two lists containing the measurements for each location were automatically analysed using Python scripts. Whilst parsing the lists, consecutive data points with exactly the same location were discarded, as it is considered that it is more likely that in between, the location has not been updated by the phone, rather than the location service providing exactly the same location twice in a row. Table 6.3 shows the number of remaining measurements after this elimination.

Location	P_1	P_2
N	275	239

Table 6.3: Number of data points per location after eliminating consecutive identical locations

To gain a first impression of the distribution, all points were plotted on a scatter plot. This distribution is shown in Figure 6.6. As can be seen, the data for the location P_1 is systematically out to the west of the measuring position, forming a conic pattern. Whilst most of the measurement points are quite close, there are several points that are multiple meters away. For location P_2, the large majority of the measurement points is very close to the location, off to the south of the latter. One single data point is several hundreds meters off to the northeast. Already, it can be seen quite clearly that the evidentiary location is in the midst of P_1-data points, suggesting that an LR favouring P_1 will be obtained and that the location-evidence obtained will indeed be pertinent for the question of device location. As described in Section 6.4, the angle and distance are analysed separately to address the influence of the angular bias that can be observed particularly

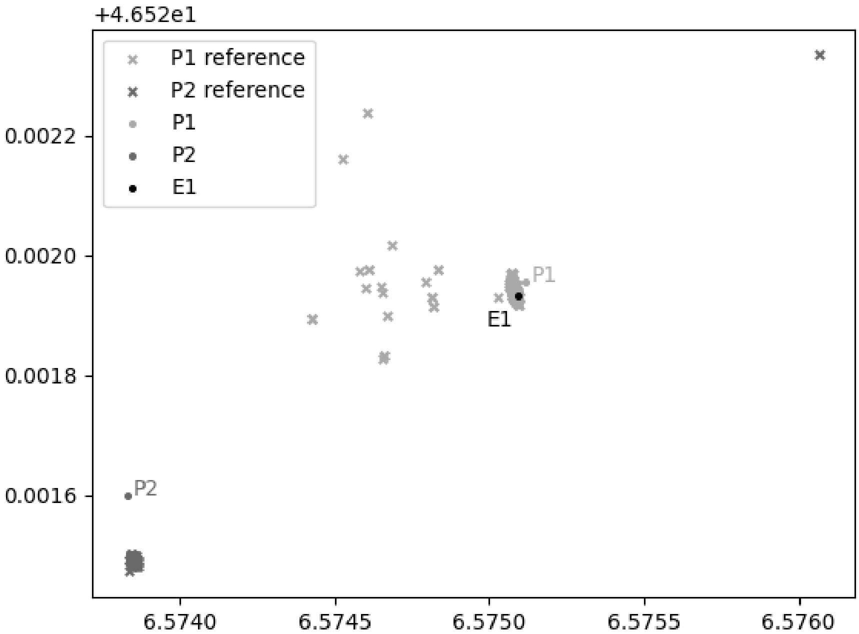


Figure 6.6: Plotted coordinates of the reference data for both locations, E_1, P_1 and P_2. (The (lat,long)-values are directly used as coordinates. x and x distances do therefore not correspond to reality)

well in the data for P_1. This is achieved by calculating the distance and the azimuth between the evidentiary coordinates and the relevant proposition. The values obtained are shown in Table 6.4. The same was then done for reference data point for both propositions. The probability of the angle was approximated by the fraction of data points that were within the 60° wedge where the angle of the evidentiary data is in the middle. The probability of the distance was assigned based on a distribution fitted over the measured distances between the origin and each data point for all points that lie within the wedge. The value was obtained from the function density at the observed distance.

Location	d [m]	φ [radians]
P_1	3.6	3.8851
P_2	144.6	0.2586

Table 6.4: Distance and angle observed for each position

6.6.2 Fingermark

In Photoshop, the picture of the finger was cropped to only show the finger and realigned to be upright. The image was then turned black and white and a curve filter was applied to improve the contrast between the ridges and the valleys. Finally, to correspond to the print it will be compared to, the image was flipped along the vertical axis. The finger visible in the picture presents a general pattern classified as accidental with an outer tracing ridge according to the NCIC classification (United States Department of Justice, and Federal Bureau of Investigation, 1984). A scan of an ink print of the person of interest was obtained as reference material. From the reference print, the size of the evidentiary mark was estimated based on the distance between the centre of the whorl and a minutia with a very characteristic form to the left of the right delta. This approach does bring along the issue that it only is consistent if the finger actually comes from the suspected source and may reinforce the LR in this regard. Additionally, the curvature of the finger in the picture will generate additional differences to a comparison to an imprinted trace. For the given case, it is however assumed, that the quantity of available corresponding features (see below), would be sufficiently dominant in the creation of the LR that this assumption will not impact the comparison too much. Both the reference print and the trace were loaded into the ESC-internal instance of the fingermark annotation and comparison tool PiAnoS (Furrer et al., 2020). In PiAnoS, first the trace then the comparison were

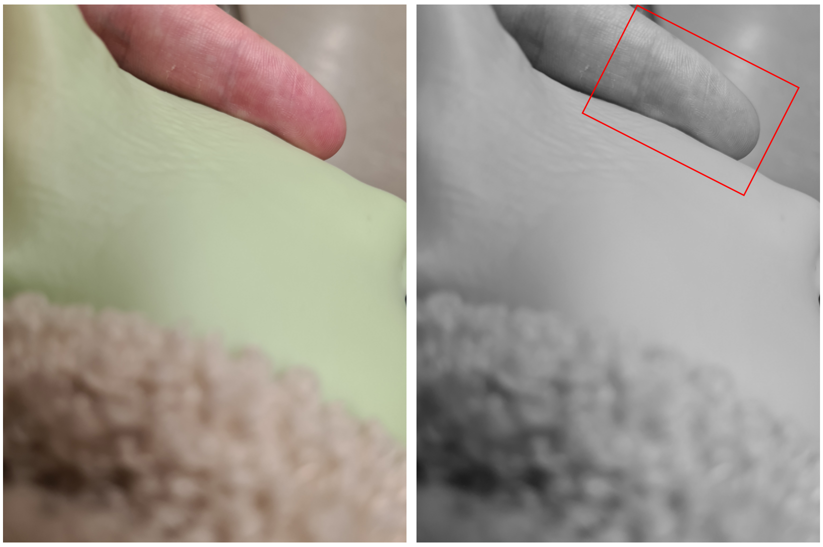


Figure 6.7: Original image recovered from the device of interest (left) and image in black and white with the area chosen for further analysis indicated in red (right).

annotated, the corresponding minutiae linked and evaluated using the built-in comparison tool. On the trace, 35 minutiae were identified, 33 of them could then be linked to minutiae on the reference print. The two remaining minutiae are located in an area of low picture quality for the trace and many different minutiae on the reference. It was consequently not clear which of the observed features corresponded and no matching was done. Several minutiae are identified on the print in regions also visible on the trace. However, the ridges in these regions are almost indistinguishable, explaining why they could not be observed on the trace. Figure 6.7 shows the original picture and the selected area. Figure 6.8 shows the treated selection and the reference print, both with and without annotation.

6.7 Results

In this section, the probability values for each piece of evidence under each proposition are discussed before combining them in an overall LR.

6.7.1 Location-Evidence

174 out of 275 points were located within the wedge for P_1 and 1 out of 239 within the one for P_2. For the distance-values, a t-distribution could be

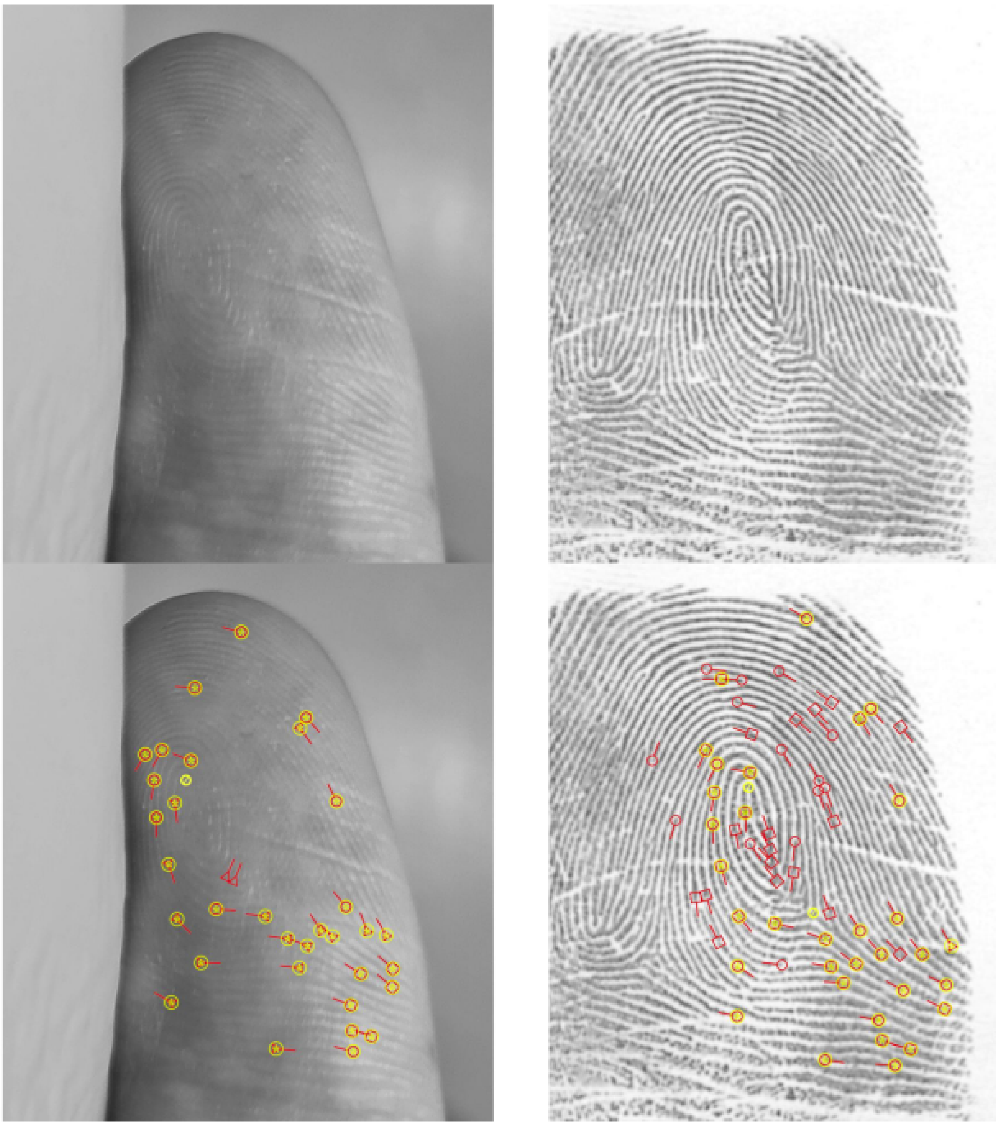


Figure 6.8: Evidentiary image, cropped, realigned and treated (left) and reference print from Person a (right); both the original picture (top) and the annotated (bottom). *Minutiae* are indicated in red, correspondences between mark and print are shown in yellow.

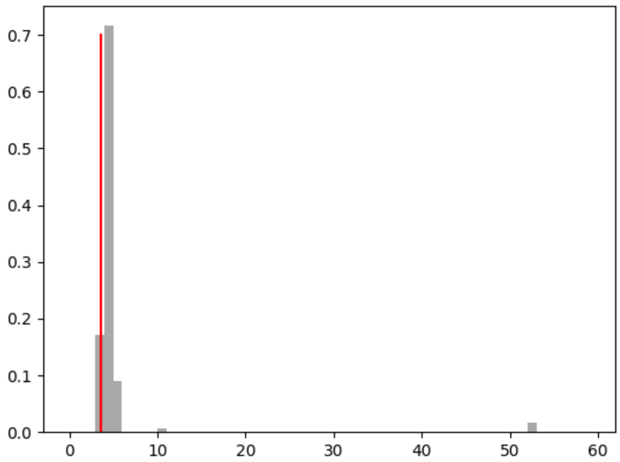


Figure 6.9: Distribution of distances within the wedge for P_1. The value observed for E_1 is indicated in red.

fitted over the values created at P_1. Figure 6.9 shows the distribution of the distances for the measurements within that wedge.

For P_2, a single data point was within the wedge, which causes a particular challenge regarding the distance-analysis for P_2 that is likely to occur in such analyses: As the evidence was created supposing P_1 to be true, there are only few data points that are within the wedge observed for the evidence under P_2, leaving behind a very meagre data set to work with for the distance-distribution. Several approaches could be envisioned for this situation:

- Instead of only the data in the wedge, the entirety of the available data is used. In the current situation, this would not create too much of an issue, as the bulk of the measurements is very close to the proposition-coordinates. In a situation however where a distribution such as with P_1 is observed, this would artificially increase the probability for the observed data in an unjustified manner.

- The small number of data points is ignored and a distribution is blindly fitted over the few available points.

- The coordinates for the proposition can be moved to the centre of gravity of the reference data. This way, some of the core data points are moved in the wedge without the more distant data points having too much of an influence on the actual distribution.

- Instead of a fitted distribution, a lower bounds probability is used. This value can be informed by how many data points are further out than the observed distance.

There is no obvious answer to the problem, none of them are easily defensible from a theoretical point of view. In the first approach, the angular dependency, which can quite clearly be shown to exist empirically, is ignored. This can lead to arbitrarily inflated probabilities. The second approach has the issue that distribution fitted over a small number of measurements are heavily dependent on singular data points and become quite volatile. If a single data point is added or removed, a completely different result may be observed. As such, it is very hard to justify an approach like this. The third approach is problematic as the proposition is adapted based on the measurements. Whilst in many cases, this may pragmatically be defensible, as the coordinates assigned to the proposition may not quite be as precise and the centre of gravity of the data cloud is often quite close to the point that was assigned, it is not a sign of good scientific practice. It becomes particularly problematic, if the points have been agreed upon by the parties, as the expert would be clearly overstepping his boundaries. The final approach has the issue that it is not evident to find a good value as a lower bound, it is however less problematic to defend from a conceptual point of view. Recognizing that the ideal approach leads to unstable results in the scenario at hand, an expert would assign a value based on his experience. This is the approach recommended by the author and applied here.

The value to choose is the probability to observe a value at a distance of 144m, given that the device was indeed at position P_2 and said location is in the wedge northeast of P_2. The single point that was observed within the wedge was at 260m from position P_2, which is almost double the distance of the evidence. Nevertheless, it would be expected that locations are closer to the actual location than what was observed. The value of 0.05 is assigned here as a very conservative value where the risk of blowing up the final LR too much is minimal, given that the $Pr(d_1 \mid \varphi_1, P_1)$ is only double the probability despite it being in the bulk of the reference. The probability values obtained from the analysis are summed up in Table 6.5.

With these results, it is possible to calculate a device-level LR based on Formula 6.3.

$$LR_{E_1} = \frac{Pr(d_1 \mid \varphi_1; LocD_1)Pr(\varphi_1 \mid LocD_1)}{Pr(d_2 \mid \varphi_2; LocD_2)Pr(\varphi_2 \mid LocD_2)} = \frac{0.640 * 0.116}{0.004 * 0.05} = 371.2 \quad (6.4)$$

A device-level LR of 371 is obtained. This corresponds to the maximal value that may be obtained for the person-level LR.

Location	$Pr(\varphi \mid P)$	$Pr(d \mid \varphi; P)$
$LocD_1$	0.640	0.116
$LocD_2$	0.004	0.05 (0.000*)

Table 6.5: Probabilities assigned based on the analysis.
*: As described, the value for $Pr(d_2 \mid \varphi_2; P_2)$ was assigned based on the experts personal knowledge and experience. The value provided by the model is indicated in brackets.

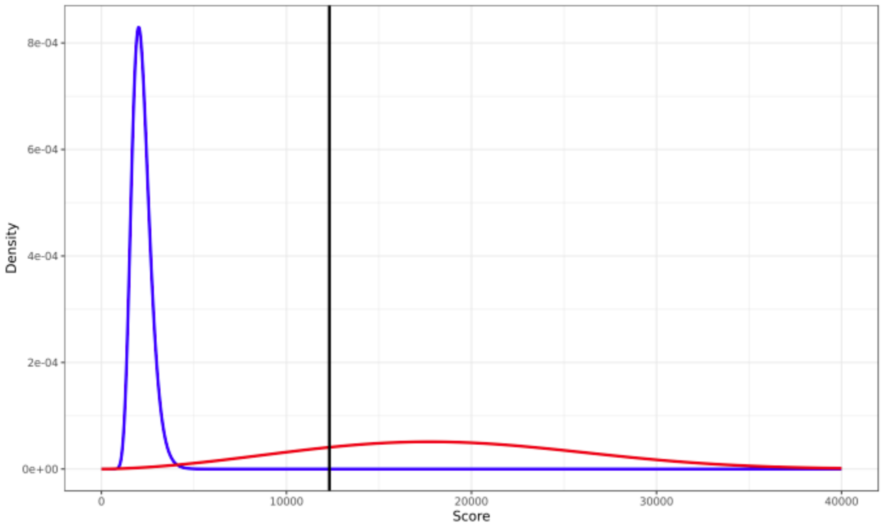


Figure 6.10: Score distribution under the assumption of same source (red) and different source (blue). The black bar indicates the score obtained from the evidentiary image compared to the print.

6.7.2 Fingermark

The LR-module of PiAnoS returns for the annotated minutiae a calibrated LR of 3×10^{10}. Figure 6.10 shows the density functions obtained from the automated evaluation process. Table 6.6 shows the probability table of E_2 based on this assessment. The probability of E_2 given $UseP_3$, the case where no one was using the phone at this moment in time is assigned at 0, as it would be quite hard to explain how this picture was created if no one was using the phone at the moment of interest.

The obtained values signify a very high support for usage at the moment of interest and it is to be expected that the overall LR is therefore very close to the LR on the device-level. To observe the effect of the considered possession, a thought experiment is conducted, in which only the general pattern is taken

$UseP$	$UseP_1$	$UseP_2$	$UseP_3$
$Pr(E_2 \mid UseP_n)$	4×10^{-5}	5×10^{-17}	0
$Pr(\overline{E_2} \mid UseP_n)$	$1 - 4 \times 10^{-5}$	~ 1	1

Table 6.6: Conditional probability Table of node E_2 based on minutiae comparison.

into account. As indicated in the previous section, the general pattern is an accidental with external ridge tracing(Corresponding to the NCIC code of XO). If the suspect is indeed the person at the source of the picture, the general picture is expected to correspond and the probability is therefore 1. In Champod et al. (2017), frequency data from an FBI collection is published showing the frequency of each category of general pattern for each finger. If it is assumed that it is known that the finger in the picture is a left index, the probability to observe the general pattern of XO is 0.003. The assumption about the finger is considered to be defensible, as it not only corresponds to the most natural position of the hand, it is also the most favorable to the defendant, as the XO pattern is the most likely to be observed on the left index. The probability table for this situation is shown in Table 6.7. This configuration gives an LR on the question of possession of 333.

$UseP$	$UseP_1$	$UseP_2$	$UseP_3$
$Pr(E_2 \mid UseP_n)$	1	0.003	0
$Pr(\overline{E_2} \mid UseP_n)$	0	0.997	1

Table 6.7: Probability Table of node E_2 based on the general pattern.

6.7.3 Overall LR

The overall LR is obtained by inputting the values shown in this section so far into the Bayesian Network shown in Figure 6.4. As expected, using the values obtained from the minutiae comparison, the LR on the person-level is 371, which corresponds to the device-level LR. According to the verbal scale proposed in (Marquis et al., 2016), this constitutes strong support for the proposition that Person A was at Location X at the moment of interest.

Using only the general pattern, the LR is slightly lower at 325. This is only marginally lower and the qualification according to the verbal scale does not change.

6.8 Discussion

For the present scenario, an LR of 371 was obtained for a full fingermark analysis and an LR of 325 was obtained with just the general pattern. The reference measurements conducted at both locations indicated that a result in favour of P_1 was to be expected, as the evidence is in the middle of the reference data points taken at Location X and the picture was indeed taken there. As such, the obtained result is what was expected.

As can be seen with the example where only the general pattern of the fingermark was considered, even when the support for the proposition of possession is not astronomical, an LR close to the device level can be obtained. It can therefore be concluded, that considering the person-level is not very impactful if direct evidence is available.

This chapter not only presents an approach on using finger friction ridge patterns to mount to person-level propositions, it also presents an approach to evaluate GPS-evidence with proportional effort.

Experiments treating the probability function for $Pr(E_1 \mid P_n)$ as a two dimensional function of d and φ should be considered as a possibility for further research into the probabilistic treatment of localisation evidence. This would eliminate the requirement to find an adapted ε value for the experiment, is however likely to require a larger number of data points to obtain a stable probability distribution. A larger scale study should also be conducted to investigate the universality of the present approach and to properly calibrate the LR. Finally, for future reproductions of the here presented approach, it should be considered to automate the registering of reference measurements.

Chapter 7

Scenario 4

In this chapter, a scenario with a location-focused question involving indirect evidence is presented. Three pieces of evidence are evaluated:

- Evidence of general usage
- Evidence that the general user utilized the device at moment of interest
- Evidence of location at the moment of interest

These three pieces of evidence are evaluated all together under a pair of person-level propositions.

7.1 Description of Scenario

This scenario illustrates the use of indirect evidence of ownership when evaluating digital traces under person-level propositions. In that setting, the usage of the device is not just contested for a given moment, but for the device in general. Evidence for all three sub-stages (general usage, usage at time t and location) is required. The following scenario illustrates a situation where this would be the case.

> A crime was committed at Location X and time t and a mobile device is found in close proximity to the crime scene. A suspect, Person A, is arrested some time later. The suspect denies all involvement with the crime and claims never to have seen the found device. When searching the apartment of the suspect, a note with a password is found that allows to unlock the found device.

In this Scenario, the following pieces of evidence are observed:

- Password
- Behavioural Pattern
- Location found within the Device

These pieces of evidence are evaluated under a pair of location-focused, person-level propositions with a specific location as the alternate proposition:

P_1:	Person A was at Location X at time t.
P_2:	Person A was at Location Y at time t.

7.2 Theoretical Background of Passwords

Passwords are a typical example of a "Something you know"-identifier as specified in Section 2.3. They have established themselves as a de facto standard for user authentication on a large multitude of devices and accounts. They have although also been criticised as an insecure means of authentication, as experience has shown two major issues: first, people tend to re-use their passwords over multiple sites, creating a security risk if the password for a given site is compromised (Ives et al., 2004; Das et al., 2014). This behaviour has also frequently been used by investigators and forensic practitioners. If a suspect refuses to give up his or her password, passwords that are used by that same person for other accounts are a valuable piece of information as they may have the potential to allow access to the account or device of interest as well. If password reuse is sufficiently persistent behaviour to allow investigators access locked devices, the suggestion can be made that reused passwords could be used as a means of identification. If it is indeed possible to access the account of a given person by entering a password that is known to be used by this person on another platform, the conclusion that the person may also be the user of this device seems to impose itself.

If password reuse is considered as evidence in a Bayesian manner, two factors of the issue need to be known: How likely is it, that a same person uses for the device or account of interest a password he or she already uses elsewhere (intra-variability), and how likely it is that another person has used the very same password independently of the first person (inter-variability). These two factors are discussed in the following.

Intra-variability: The ideal source to gain information about the reuse of passwords by a given person would consist this persons password manager if the person uses one and access can be obtained. This will not always be the case as password managers are still only used by a, admittedly growing, minority of persons (Gaw and Felten, 2006; Google and Harris Poll, 2019). Over the years, a series of studies based on questionnaires have been conducted to assess the behaviour of persons when using passwords (Dhamija and Perrig, 2000; Brown et al., 2004; Riley, 2006; Gaw and Felten, 2006). Albeit generally very thorough, these studies are limited by their relatively small sample size and being limited to a specific population, either students (Brown et al., 2004; Riley, 2006; Gaw and Felten, 2006) or employees of a given firm (Dhamija and Perrig, 2000). A very large scale study was conducted by Florencio and Herley using a optional extension for the Windows Live toolbar. Their measurement method consists in recording inputs in fields designated as password-entry fields (Florencio and Herley, 2007). Their method has the advantage of measuring password behaviour at a large scale and independent of user self-reporting has a huge advantage of making their results more objective as it is not dependent on users memories and honesty. However, as the authors are not able to distinguish between erroneous password entries and actual password entries, they decided to eliminate all entries within their recorded data that have a complexity below 20 bit complexity are ignored. The authors consider this to be just a minor source of error. However, analysis of the real world password dump from Burnett (2015c) indicates that this assessment is likely wrong. Indeed, this criteria eliminates about 3% of the 1'000 most frequent password in the data set, indicating a large quantity of discarded passwords. In recent years, some vendors of password management tools have conducted analyses about the password habits of their customers and published them for mostly promotional purposes (c.f. LastPass, 2019). These studies are focused on users of password management services, likely reporting better behaviour than for a wider population, as users of a password manager are by default more likely to be aware of risks related to passwords. Nevertheless, these studies do give an interesting insight in the habit of users regarding passwords. As they are based on actual passwords stored and can refer to a large quantity of data, this study is likely to be representative for this specific population. Table 7.1 gives an overview of reported password numbers per person and number of unique passwords per persons, allowing to draw conclusions about password reuse.

Inter-variability: A widely known fact about passwords is that there is a very low diversity in the passwords that are chosen by different users. In-

Study	Year	N	# of PW	# of unique PW
Dhamija and Perrig	2000	30	10-50	1-7
Brown et al.	2004	218	3-20	1-11
Riley	2006	328	-	3.1*
Gaw and Felten	2006	49	7.8*	3.3*
Florencio and Herley	2007	544'960	25*	8.11*
LastPass	2019	47'000**	75	5***

Table 7.1: Reported rates of password reuse.
(*: Mean values reported; **: Number of organisations ranging from 1 to 10'000+ employees; ***: calculated based on reported average password reuses)

deed, based on collections of compromised credentials is has been estimated that the 1'000 most frequent passwords allow to access 85% of all online accounts. These collections allow to estimate how frequent a given password is, although there are strong indications that these dumps may not always be well adapted for this task. As generally, services with lower security standards get leaked, it is well possible that many of the passwords were chosen with a lower requirement for security than would have been chosen for more important services. Additionally, it has been suggested, that requirements by the service have a major influence on the passwords chosen. (Florencio and Herley, 2007). It can therefore be expected that the requirements of the password will have an influence on the choice of password. Similarly, the type of input very likely has an influence and passwords typically entered through a computer keyboard will not have the same characteristics as passwords entered from a smartphone. For example passwords with letters all on one row of the keyboard are generally more frequent (Brown et al., 2004). As these arrangements change from one region to another, so are password frequencies likely to vary.

A study by Kanta et al. recently found that the type of service for which the password is used will have an influence on the password as well. For example, a wordlist generated from a Manga forum password dump will be more efficient in cracking passwords for Manga-related services than other passwords list are (Kanta et al., 2021). There is also a high probability that the choice of password is influenced by the language spoken by the person choosing the password, at least for passwords containing words.

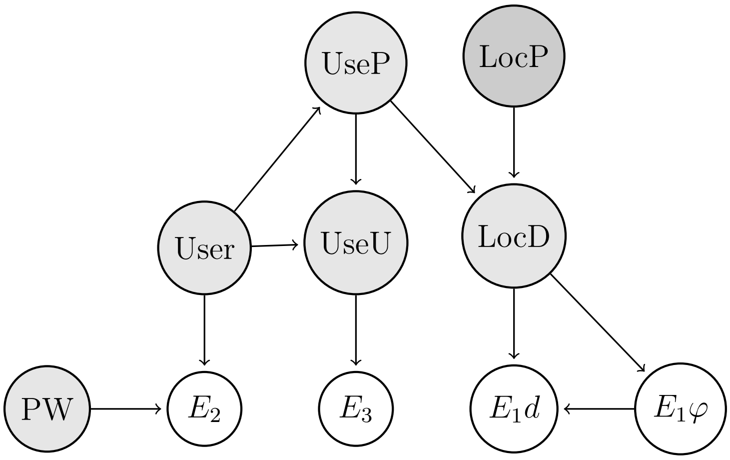


Figure 7.1: Bayesian Network for scenario 4. As in scenario 3, the $E1$-node has been split up to take into account the distance and angle of the evidence. The node PW has been added to simplify the evaluation of E_2.

7.3 Discussion of the Framework

The Bayes Net used in this scenario (cf. Figure 7.1) is based on the network shown in Figure 3.2 in Chapter 3. The node PW is adapted to model three different possible behaviours of Person A regarding password reuse: Person A never reuses a password (PW_1), Person A sometimes reuses a password (PW_2) and person A only uses one password for all their services (PW_3). Table 7.2 shows the probability table of node E_2, the password evidence, adapted accordingly. Given that node PW only describes Person A, its state only has an impact on the probabilities of E_2 if Person A is the general user of the device, in other words, if $User_1$ is true. In this case, if Person A never reuses a password (PW_1), they would not have chosen the password found at their apartment for their mobile device as well and the probability to observe E_2 is therefore 0. If Person A uses the same password everywhere (PW_3, the probability to observe E_2 would become 1, as they would certainly be using this password. Finally, if they sometimes reuse their passwords (PW_2), there needs to be an assessment made about the probability of Person A using this particular password for the phone. This probability (f_1) together with the probability of E_2 if $User_2$ is true and the prior probabilities in node PW are discussed in Section 7.5.3.

$User$	$User_1$			$User_2$			$User_3$		
PW	PW_1	PW_2	PW_3	PW_1	PW_2	PW_3	PW_1	PW_2	PW_3
E_2	0	f_1	1	f_2	f_2	f_2	0	0	0
$\overline{E_2}$	1	1-f_1	0	1-f_2	1-f_2	1-f_2	1	1	1

Table 7.2: Probability table of node E_2.

7.4 Simulation of Data

All data used in this scenario was generated on an iPhone 6s (A1688) running under iOS 14.4.4, although not at the same time. The behavioural data used is the same as in scenario 2 and the location data was created simultaneously to the data used in scenario 3. As such, the two locations X and Y remain the same as in scenario 3. Their coordinates are shown in Table 7.3.

	Longitude	Latitude
P_1: Location X	6.575116326	46.521954786
P_2: Location Y	6.573832039	46.521592273

Table 7.3: Coordinates of the positions considered in each proposition.

7.4.1 Localisation

As with scenario 3, pictures were taken as the phone was sitting on the table at both locations. The period during which the pictures were taken was somewhat shorter and only 699 pictures were obtained. The quite important difference to the number obtained in scenario 3 is mostly due to the realisation that a longer interval between two pictures is required for a location change to actually take place, and so the rhythm at which pictures were taken was reduced. A logical extraction of the device was conducted with Cellebrite UFED 7.53.0.24 extracting only pictures. The result was opened in Cellebrite Physical Analyser 7.54.1.7 and an Excel-report was generated from the locations tab. Manually, this report was split up into two Excel lists, each one containing the reference data points for a given location.

7.4.2 Password

A password has to be chosen as the password found by the investigators. During the analysis stage, this password's frequency will be compared against the frequency of the password in a reference sample. To observe the influence of password rarity, three different passwords are chosen as different version

of the evidence: The most frequent password (*123456*), a password with medium frequency, in the dump ranked at 1'000, (*wildcat*) and a password that is not present in the dump (*dorigny*).

7.4.3 Behavioural Biometrics

As the behaviour biometrics-evidence, the data from scenario 2 was reused. In this scenario, the alternative hypothesis is an open set population and not a specific person, the data for the alternative population is expanded by the anonymised data used in Michelet (2021). This data consists of a set of 7 different persons, using four different devices of the same make and model used here. The data set contains more than just the system characteristics recovered from the devices in Chapter 5. These additional characteristics are filtered out and the remaining characteristics are integrated in the group of reference data under proposition P_2.

7.4.4 Considerations for a Real World Case

The approach to assess the probability of password reuse attempts to construct a general probability that *any* person, randomly chosen from the general population, would reuse a password. This is likely not particularly well adapted to a specific person, as password reuse is highly dependent on awareness, background and other general habits of the person of interest. Here, a general probability of any person belonging to one of three groups of behaviour is used to construct the above probability. Based on the available information of a person, it is likely possible to know which group the person actually belongs to. Often several passwords used by the person on multiple accounts are known to investigators, giving an insight into password behaviours of the person. If lists of passwords or password storage services are available, the likelyhood of the person reusing a given password may directly be inspired by these. Regarding the frequency of a password, it may be possible to generate a more adapted reference data set based on the data stored in case management systems of the lab. Often, passwords of analysed devices are recorded in a dedicated field when documenting the device, as this information may be required during the extraction. If the analyst can query this field in his database, he may be able to obtain a very relevant reference collection of passwords, specific to a given region, specific to a type of device and specific to the part of the general population whose device is likely to be analysed by a forensic expert.

At the moment of writing, there are no publicly available data sets of behavioural biometric reference data, and it may not be evident to obtain such a data set. If a lab envisions conducting such analysis on a regular basis, quite a bit of thought should be put into how such a dataset may be constructed with proportional effort.

7.5 Conducting the Analysis

For this scenario, the three elements of evidence are analysed separately and the results are combined only at the end. As the approach for GPS analysis is already presented in Chapter 6 and behavioural biometrics is presented in Chapter 5, these approaches are not explained in detail anymore here. Instead, the approach is broadly described, stating the specific values for the present scenario and only describing details where the process diverges from what was done in the two earlier chapters.

7.5.1 Localisation

To analyse and evaluate the Location-Evidence, the same approach was followed as in Chapter 6. The location indicated by the evidence E is shown in Table 7.4

	Longitude	Latitude
$E1$	6.573944444444444	46.5213305555556

Table 7.4: Coordinates recovered from the Evidence E_1

Again, eliminating identical consecutive coordinates, the measurements were automatically analysed using a python script. Table 7.5 shows the number of remaining measurements after this elimination. A plot of all those coordinates can be found in Figure 7.2. As was already the case in scenario 3, the measurements spread out to the west for Location X and to the south for Location Y. However, this time the spread is quite a bit larger. E_1 is right in the middle of the points measured at Location Y, suggesting an LR in favour of P_2 will be obtained. Table 7.6 Shows the observed angles and distances under each proposition.

7.5.2 Behavioural Biometrics

The analysis process of the behavioural biometric evidence is done following the same approach as presented in Section 5.5 : The data is normalised and

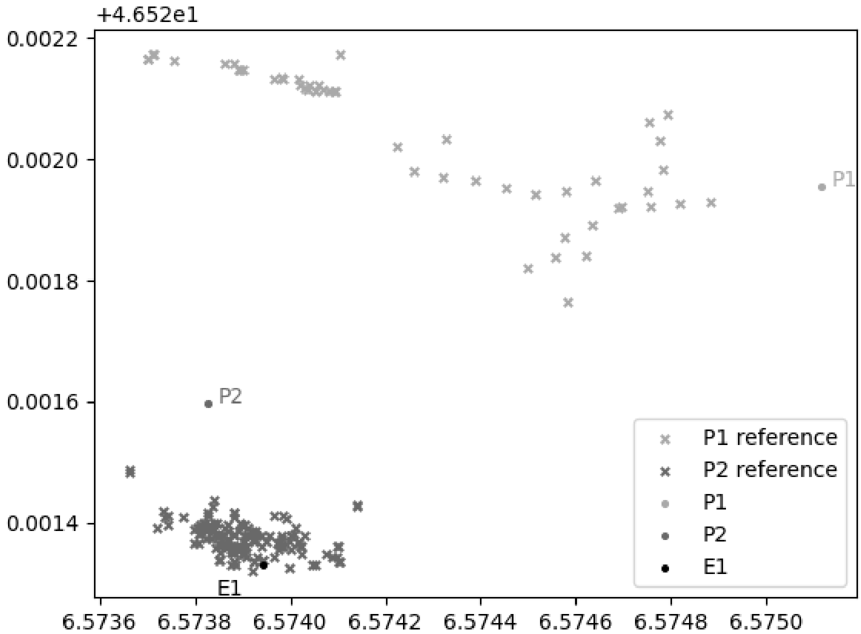


Figure 7.2: Plotted coordinates of the reference data for both locations, E, P_1 and P_2. (The (lat,long)-values are directly used as coordinates. x and y distances do therefore not correspond to reality)

Location	P_1	P_2
N	50	149

Table 7.5: Number of data points per location after eliminating consecutive identical locations

Location	d [m]	φ [radians]
P_1	146.8	3.6310
P_2	32.3	5.1235

Table 7.6: Distance and angle observed for each position

a PCA is conducted. In scenario 2, the first PC was sufficient to separate out the two distributions. This is no longer the case here. Visualising the distribution of the first 40 PC, no evident cut off point is visible (cf. Figure 7.3). A decision is made to use the first 10 PC as values. A data point for single day therefore consists of a 10-dimensional vector where each dimension is the corresponding PC-value of this day. Geometric analysis is conducted in 10-dimensional space instead of one dimensional. The distances compared for the evidence are obtained by calculating the distance of the observed data to the center of mass of the reference data. The results are shown in Table 7.7.

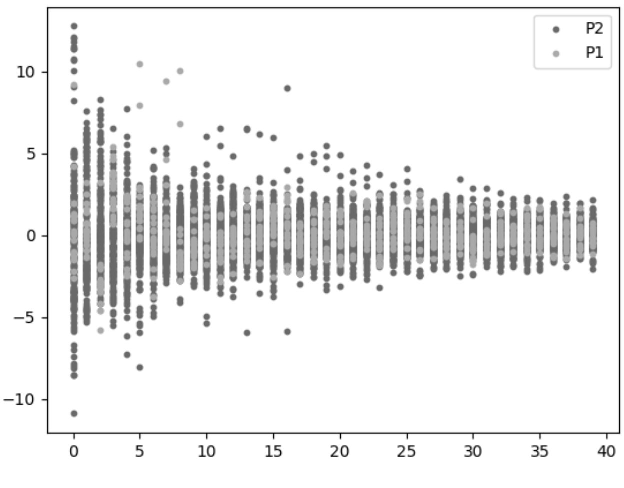


Figure 7.3: PC-value of the first 40 PC for reference values observed under P_1 and P_2

Subscenario	S1	S2
Distance	6.588	4.055

Table 7.7: Distances observed in behavioural biometric analysis for the day of interest in both subscenario.

7.5.3 Password-Evidence

To assess the impact of the password evidence, two values must be assigned: First, the probability of this password unlocking the phone if the suspect is indeed the general user of the device. Second, the probability, that the password unlocks the device despite the user not being the general user of the device.

Given the suspect as the primary user

If the suspect is indeed the primary user, the probability whether they would be reusing the password found in the apartment would depend on their behaviour regarding password reuse. As already indicated in Section 7.3, a node is added to the Bayes Net to model this behaviour. This node PW contains three states: Person A never reuses passwords (PW_1), Person A sometimes reuses passwords (PW_2) or Person A only uses one password for all their accounts (PW_3). A 2019 survey of 3000 adults in the US asked participants to answer in which of those categories they fall (Google and Harris Poll, 2019). The reported frequencies from this study are used here to approximate the prior probabilities of Person A falling in each of the presented categories. The probability table of PW is shown in Table 7.8.

PW	$Pr(PW_n)$
PW_1	0.35
PW_2	0.52
PW_3	0.13

Table 7.8: Probability table of node PW. Data from Google and Harris Poll (2019).

Based on this separation, only a value for the probability of E_2 if the primary user is Person A and Person A sometimes reuses their passwords ($Pr(E_2|User_1; PW_2)$) has to be assigned. This probability is assigned based on another study from 2019, when LastPass, a provider of password management services, published an analysis based on the data from 47'000 enterprises using their services. They found that in Switzerland, the average

person has 74 passwords stored in their application. This value is very close to the international average of 75. Additionally, the average employee, both in Switzerland and internationally, reuses a password 13 times (LastPass, 2019). If it is assumed, that all passwords are equally probable to be used by a given person, there is a probability of $^{13}/_{74}$ that an average person would be reusing a specific password. This value is assigned as the probability $Pr(E_2|User_1; PW_2)$.

Given another primary user

If someone else is the primary user, this comes down to the probability of someone else choosing the password that was found at the suspects home. It is here assumed, that whoever was the person that chose the password would not be aware that the suspect is using this password as well. Given the high degree of password reuse (cf. Section 7.2), this is not per se an outlandish possibility. To quantify this possibility, a data-set of real world, leaked passwords is used. This password dump containing 10 million passwords was published by Burnett in 2015. The dump is a collection of different password leaks, and while some passwords have been removed as they were linked to sensitive infrastructure or allowed to identify their users, it is generally quite representative of the overall distribution of data sets (Burnett, 2015c,b). At the moment of writing, the data set is not available on the original site anymore, but a copy is available on the Internet Archive (Burnett, 2015a). From this data, the required probability is assigned based on the frequency of the observed password within the password dump. The specific values are discussed in Section 7.6.2.

Several criticisms can be raised regarding the choice of the data set:

The age of the data set: The data set was published in 2015c and whilst overall, password frequencies have not massively changed over the years, some tendencies could be observed, such as «12345678» overpassing «123456» as the most frequent password, likely because an increased number of services requiring at least 8 character long passwords for their services. This effect as well as passwords containing cultural references more recent than 2015 would not be represented in this data set.

The sources of the data set: The data set is comprised of passwords recovered from password dumps, mostly when the databases of online services were compromised (Burnett, 2015b). Given that system requirements and limitations will impact the choice of password, it can rightly be argued that

the passwords are not a good representation of the passwords the population of potential passwords.

The language of the data set: The majority of the users of the data set appear to be heavily influenced by the English language. For example, the English word «monkey» is the 15th most frequent password in the data set with 3246 appearances throughout, whilst the French «singe» ranks 51'696th with just 16 appearances. If the potential user of the device is suspected to not be English speaking, a strong argument can be made against the use of this data set.

The cultural context of the data set: As has been shown in (Kanta et al., 2022), the cultural environment in which a password was chosen, impacts the choice of password. The present data set is thought to be a more or less general data set given its size and the way it was constituted. Consequently, it may not be adapted if the potential «other» users are not the general population, but a subgroup of a very specific cultural context. If the phone was for example found in a football stadium, football related terms are likely to be underrepresented in the data set.

As can be seen, the data set is by far not ideal for the analysis at hand. However, for the present proof of concept, it is considered sufficiently appropriate.

7.6 Results

This section presents results obtained for the different types of evidence. The results are presented regarding the question of device location first, device ownership second and, combining the two intermediate results to an LR on the question of the location of the person. At both intermediate stages, an LR is given on the level of the addressed question to give an impression of the strength of the observed values.

7.6.1 Location of Device

19 out of 50 points were located within the wedge for P_1 and 137 out of 149 within the one for P_2. T distributions were fitted over both distributions, as these distributions resulted in the smallest sum of square errors. Figure 7.4 shows the distribution of the distances for the measurements within both wedges.

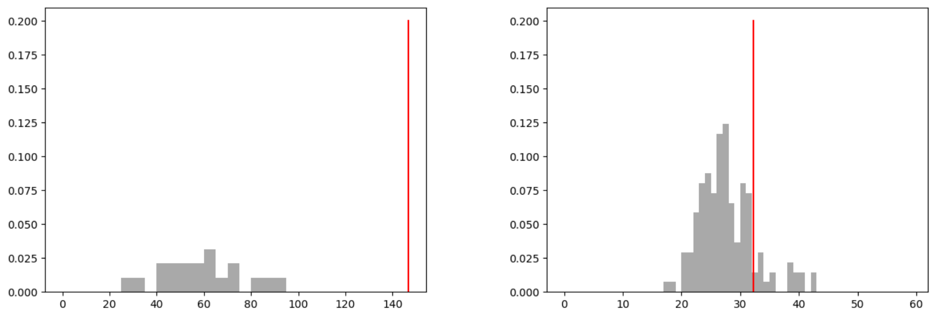


Figure 7.4: Distribution of distances within the wedge for P_1 (left, n=19) and P_2 (right, n=137). The value observed for E_1 is indicated in red.

This time, the distribution under P_1 is based on a very small number of values (19). Additionally, the observed value lies in the tail end of the distribution which is more likely to be affected by variations in the overall function. To address this issue, the choice was made to assign a lower bounds value of 0.001 to $Pr(d_1 \mid \varphi_1; P_1)$ to not assign probability values far below what is justified with the size of the reference data sample.

With these results, it is possible to calculate a device-level LR based on Formula 6.3 presented in Chapter 6.

$$LR_{E_1} = \frac{Pr(d_1 \mid \varphi_1; LocD_1)Pr(\varphi_1 \mid LocD_1)}{Pr(d_2 \mid \varphi_2; LocD_2)Pr(\varphi_2 \mid LocD_2)} = \frac{0.380 * 0.001}{0.919 * 0.054} = 0.007657 \tag{7.1}$$

A device-level LR_{E1} of 7×10^{-3} is obtained (cf. Formula 7.1). As values below one are generally not very well readable, an inversion of the proposition can be envisioned (ENFSI, 2010), where the above value correspond to an LR of 130 in favour of $LocD_2$.

Location	$Pr(\varphi \mid P)$	$Pr(d \mid \varphi; P)$
$LocD_1$	0.380	0.001 (7×10^{-8}*)
$LocD_2$	0.919	0.054

Table 7.9: Probabilities obtained from the analysis.
*: As described, the value for $Pr(d_2 \mid \varphi_2; P_2)$ was assigned based on the experts personal knowledge and experience. The value provided by the model is indicated in brackets.

7.6.2 Usage of Device

Password

The password dump contains 5’189’382 distinct passwords. Table 7.10 shows the observed number of appearances for the considered passwords. «Dorigny» does not appear in the data set. It is therefore assigned a value of 1.00×10^{-7} corresponding to it appearing one single time, as the value is considered to be at its highest that high, but there being insufficient data available justifying assigning a lower value. Inputting the values in the Bayes Net, an LR can be obtained on the question of who is the general user of the device. The obtained LR are all in favour of Person A being the general user($User1$). The strength of the support varies heavily depending on the rarity of the password, from 40 for *123456* (qualified as moderate according to Marquis et al. (2016)), over 6’687 for *wildcat* (qualified as strong support), to 2.214×10^6 for *dorigny* (qualified as extremely strong). Table 7.10 shows the assigned probabilities and obtained LR for the different passwords.

Password (E_2)	rank	# of appearances	$Pr(E_2 \mid User_2)$	LR
123456	1	55’893	5.59×10^{-3}	40
wildcat	1’000	331	3.31×10^{-5}	6687
dorigny	-	0	1.00×10^{-7}	2.214×10^6

Table 7.10: Number of appearances and of the considered passwords in the reference dump, assigned probability and obtained LR in favour of Person A being the general user of the device ($User_1$).

Behavioral biometrics

The visualisation (cf. Figure 7.5) of the observed reference values shows quite a bit of overlap between the two populations. It is therefore to be expected, that no particularly high LR values will be obtained. Plotting both observed values, it can be seen that the evidentiary value in sub-scenario 1 is very close to the overlap.The evidence in sub-scenario 2 is in the midst of the distribution from P_1. A burr12 distribution was fitted over the P_1-values and a double weibull distribution was fitted over the P_2-values. From these distributions, density values for the observed evidentiary distances were obtained. Figure 7.6 shows the density functions and Table 7.11 shows the probabilities obtained based on these distributions.

These values are fed into the Bayes Net to obtain the LR for the question at hand. They do however also allow to obtain an LR on the question of the abstract user by comparing $Pr(E_3 \mid UseU_1)$ and $Pr(E_3 \mid UseU_2)$. The

Sub-scenario	S1	S2
$Pr(E_3 \mid UseU_1)$	0.05951	0.25435
$Pr(E_3 \mid UseU_2)$	0.10816	0.02897

Table 7.11: Probabilities of observing the evidentiary distance under each proposition for both sub-scenarios.

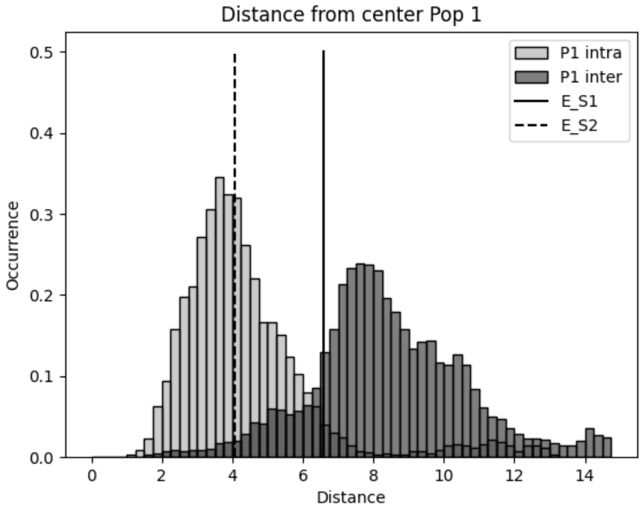


Figure 7.5: Histogram of distances of behavioural biometric observations.

obtained LR are shown in Table 7.12. These LR values are quite low, in the case of S1 even wrongly in favour of $UseU_2$, although very feebly with a value of approximately two[1].

Subscenario	S1	S2
LR on abstract User	0.5464	8.779

Table 7.12: LR in favour of $UseU_1$.

7.6.3 Location of Person

The probability values obtained in the previous sections were input into the Bayes Net and LR for each scenario were calculated. As all LR were in favour of P_2, the propositions were exchanged to obtain LR over 1. Table 7.13 shows the likelihood ratios in favour of Person A being at Location Y (P_2). All LR

[1]This value is obtained by exchanging the propositions as proposed by (ENFSI, 2010)

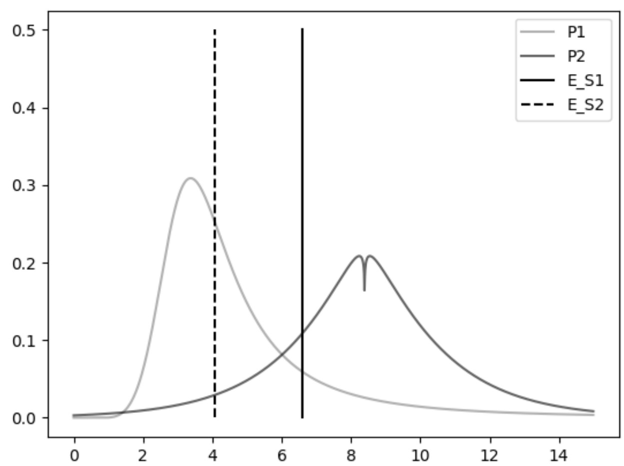


Figure 7.6: Density distribution of behavioural biometrics distances.

support the correct proposition, are however all very close to one. According to the scale by Marquis et al. (2016), the support is qualified as weak or limited. The reason for this is the very weak support from the behavioural biometrics regarding the abstract user of the device.

Behavioural Biometrics Password	S1	S2
123456	2.2	7.5
wildcat	2.2	7.6
dorigny	2.2	7.6

Table 7.13: Table containing the overall LR in favour of P_2 for each scenario

7.7 Discussion

The obtained LR are slightly in favour of P_2, but do not really add much value. The issue stems from the very low values of the general user having the device on them. As such, even though the analysis of the password heavily supporting the suspect being the general user of the device, there is almost no added value to the initially asked question. In such a situation, the expert should consider whether the evidence available is indeed adapted to express an opinion on the level of the person.

It can also be observed, that the evidence of the password does not do much to influence the final result. It's impact is minimal, the LR mostly being dominated by the behavioural biometrics result.

The absence of the password *dorigny* from the reference data set illustrates two issues:

- How to handle an absent password.
- The issue of how well the data set is adapted to the situation at hand.

Especially the second aspect is quite evident. Dorigny being the name of the campus at the University of Lausanne, most persons being present there are likely to know this word and potentially susceptible to choosing the name as a password. In a general, worldwide population, Dorigny is unlikely to be a name that is known and therefore chosen as a password[2].

In the present scenario, a password without a particular relation to the person was considered. It is known from research that sometimes, people tend to include identifying information in their password, such as their name or birthday (Brown et al., 2004). If such a password were to be found, a more complex evaluation approach could be considered based on the frequency of the characteristic used in the password.

[2]It may even be argued, that this fact questions the independence between E_2 and E_1

Chapter 8

Discussion and Conclusion

Location-related evidence is a valuable asset for criminal courts, as for most crimes, a person has to be on the crime scene to commit the act. Whilst having been used extensively for quite some time, and due to wider availability also increasingly more often, there has been put only little thought into their reliability. A wide variety of sources for uncertainty exist for digital location-related traces, some linked to the way they are created and analysed, others a direct consequence of the person-device gap, an issue inherent to the problem at hand. These uncertainties can only be addressed by properly evaluating the observed traces. However, formal evaluation of digital evidence in general is to this day not widespread among practitioners, and when it is done, it is rarely approached in a structured manner.

In this work, a structured approach to address uncertainties linked to location-related evidence was presented. This approach is flexible, allowing to address situations of device- or person-level. Such an approach is urgently needed, as practitioners become more aware of limitations and are looking for solutions (Bassi and Scoundrianos, 2022), courts reject presented evidence (Poser, 2017; Swiss Federal Court, 2019) and emerging standards call for the application of a logical framework (ISO/TC 272 Forensic sciences).

8.1 Signification of Scenario Results

Two research hypotheses were presented in the beginning of this work. Hypothesis 1 was stated as follows:

Research Hypothesis 1. *It is possible to gain, from a mobile device, for a given moment in time, relevant traces about where that device was and who was using it, allowing an expert to express an opinion on a person's whereabouts, supported by a structured reasoning process.*

Whilst not completely surprising, this work has shown that such traces can exist on a mobile device. Already in existing literature, these possibilities are discussed (cf. Chapter 2), but the results in this thesis show that evidence capable of distinguishing between propositions of the discussed type do clearly exist. Chapters 4, 6 and 7 present scenarios where evidence differs depending on whether it was created at one location compared to another. Chapters 5, 6 and 7 present scenarios with evidence that differs from one user to another. As such, the here presented results confirm the stated research hypothesis.

This finding is unlikely to be surprising to specialists in the field, however, to the knowledge of the author, this reasoning has been explicitly stated for the first time in this work.

Hypothesis 2 was stated as follows:

Research Hypothesis 2. *Traces from a mobile device can be evaluated in a logically consistent manner under a pair of location-focused propositions with a physical person as a subject.*

This work presents an LR approach for the evaluation of location-related evidence under location-focused propositions with a person-level subject. The approach is modeled as a Bayesian Network allowing for the combination of the different sub-stages of the reasoning and the different elements of evidence. As shown in Chapter 3, this Bayesian Network behaves as expected in extreme cases and its behaviour otherwise is consistent with expectations. In Chapters 4, 5, 6 and 7, four scenarios have been presented where evidence was evaluated using the proposed approach. The probabilities assigned were based on simulated data, not only contradicting the claim by Horsman (2020) that probabilistic evaluation of digital evidence is unfeasible, but also showing that the evaluation of digital traces can be done using the same approaches well established for other traces. The results obtained from the simulations were consistent with expectations, although in some cases close to being irrelevant. Overall, the work in this work confirms both the presented hypotheses.

In this work, an approach has been presented to bridge the person-device gap in relation to location-focused questions. Using this approach may allow an expert to express an opinion on the location of a specific person instead of the location of a device. This may be a relevant tool depending on the case at hand. The person-device gap is not just an issue in cases where the location of a person is questioned. Especially regarding actions committed on digital devices, this issue arises frequently. Due to the often cited large

quantity of logs, it is often not too hard to reconstruct the activity that took place. As a consequence, the question is more likely to be *«Who has used the device at this moment in time?»* than *«What activity has taken place?»*. Whilst not explicitly discussed, the approach for bridging the gap in this work is applicable to these scenarios as well. Indeed, if it can be shown, that the device was being used by a given person at time t, activities conducted by the user of that device have to be done by that person. The same logic applies for probabilistic reasoning, leaving open the possibility that the conclusion may err on the user at time t.

Whilst not explicitly an aim of the work, this thesis has yielded some additional results that are of interest.

- Scenario 1 presented in Chapter 4 gives a very rudimentary way to evaluate cell tower connections[1].
- Scenario 2 and 4 present an approach to evaluate results from behavioural profiling such as presented by Michelet (2021) and Guido et al. (2016). Whilst still requiring much more research, these approaches may be of interest in the future.
- Scenario 3 presents a novel approach to evaluate GPS-traces. This approach seems to be a promising method for evaluating this type of trace, although more testing and calibration is required.
- Scenario 3 also presents a combination of physical and digital traces to form an overall conclusion. Whilst having been postulated in (Spichiger, 2021), to the knowledge of the author, this is the first time this has been done in published literature. This enforces once more the idea that digital forensic science should be considered as a subdomain of forensic science and that digital traces should not be treated as fundamentally different from physical traces.

8.2 Application in Real World Cases

The scenarios presented in this work were simulated and chosen in a way to explain and illustrate the working of the here presented framework. In real world cases, a series of additional challenges are to be expected, some operational, some technical. This section discusses some of them.

[1] This approach is very likely to be surpassed by more advanced approaches currently being researched (cf. Bosma (2022)).

Choice of evidence: In many situations, there will not be a single observation but a multitude of data points. This work does not have a solution for that. One may be tempted to evaluate multiple observations separately and then multiply them. However, multiple pieces of location-related evidence recovered from the same device are unlikely to be independent from each other, making this approach unsubstantiated. Two approaches are possible to resolve this issue: Either, select one single observation and evaluate this one, or take a holistic approach and assign a value to the entirety of the observed data.

Time-Constraints: In criminal investigations, providing results is often time-sensitive. Results that arrive later than the moment of the decision are of no value at all (unless for a potential appeal). Given that practitioners working in digital forensic units frequently report large backlogs, keeping analysis as quick as possible is an understandable wish. The here presented approaches are all rather time intensive. For a single point of data, two days were invested at least to gather data. It is not feasible to conduct such analysis for every single location data point that crosses the desk of a digital forensic practitioner. In a real world case, the practitioner will have to work with a selection of the traces that will become evidence and do the evaluative work for those only. An approach will have to be chosen based on the time the practitioner can justifiably invest on the evidence based on the importance of the case and the present workload.

Technical Capabilities: The tools available to an entity tasked with conducting forensic analysis on a mobile device may vary heavily, mostly dependent on budgetary factors. While some police forces and forensic labs invested heavily into the development of digital forensic capabilities, other still not have advanced beyond recovering files from storage devices. Depending on the source device, and the state in which it is recovered, automated tools for extraction may not be able to recover the traces here discussed, even if these tools are available. In such situations, labs may need to request help from other organisations that have the capabilities to recover these traces.

Skill and Knowledge of the Practitioner: Not every practitioner may have the necessary knowledge to conduct the here presented evaluations. Indeed, due to a lack of specialists, personnel working in digital forensic laboratories and units often have only limited specialised knowledge on the workings of the systems of interest. Even if they have technical knowledge, they may have never been taught about how to conduct an evaluation as

presented in this work, as the concept of evaluation in digital forensic science is still very young (cf Chapter 2.1.1). A unit may find itself faced with a situation where evaluation is required without having members with the required knowledge to conduct the analysis. In such situations, it is essential that they look for help, either by tasking an expert with conducting the evaluation or by requesting an expert to assist them in conducting the analysis themselves[2].

Legal Constraints: Depending on the jurisdiction, requirements towards expert testimony may vary heavily. Due to the structure and traditions of a legal system, what the expert is allowed to express opinions on and what information is available to the expert may be different. This can have an impact on what analyses and evaluations can be conducted, on what level an opinion can be expressed and with what data this opinion can be supported. It is therefore to be expected that the strength and type of an evaluated result may vary from one jurisdiction to another. It is strongly advised to a reader interested to apply the here presented approach, to inform themselves about the legal requirements of the jurisdiction their testimony takes place in.

Understanding of Partners: For most humans, working with uncertainty is not something intuitive. As a consequence, it may be challenging for police officers, lawyers, judges and members of juries to understand the results presented in the manner here proposed. Especially in jurisdictions where a Bayesian presentation of results is not already the norm for physical traces, an expert may have to do an important bit of explaining the approach to his partners. This is not easy to do, however preliminary results of recent research suggests that reports presenting LR are equally well understood as more categorical reports (Salonen, 2022).

8.3 Future Work

This work took first steps in a multitude of young and developing domains. As such, a broad range of future research possibilities have opened up.

[2] This is somewhat of a *«teach a man to fish»*-situation. Long term, the aim must be that all entities working with that kind of trace are capable to evaluate their results in a structured manner.

8.3.1 Preliminary Phase

The reasoning structure presented in Section 1.2 begins with a preliminary stage, in which the fundamental adaptability of the traces in view of the question is investigated. In this work, it is presumed that the examiner has sufficient contextual information to verify that the traces were

- Correctly recovered and represented by the analytic tool.
- Generated on the device they were recovered from.
- Generated at the moment in time indicated by the timestamp.

It is not always possible or justified to make these affirmations in a categorical manner. An overall expression of uncertainty should consider this as well. Consequently, further research should explore the possibility and find ways to integrate uncertainties from the preliminary stage into the overall LR. This consists not only in finding ways to quantify the uncertainties linked to those preliminary questions, but also finding ways to integrate them into the overarching evaluation. Existing work on Lab error (e.g. Thompson et al. (2003)) should provide a relevant foundation to resolve this issue.

8.3.2 Cell Towers

As discussed in Chapter 4, there is currently little empiric research into the factors that influence cell tower connections of a mobile device. When looking to simulate a situation comparable to the moment of creation of the trace for reference data gathering, it is currently unclear what factors need to be addressed. In particular, systematic research into the influence of weather, seasons, the time of day and the day of the week should be conducted. For older cases, when the traces were created several years or even decades ago, a longitudinal study investigating the persistence of cell tower systems would be of interest.

A research group at the NFI is currently working at a probabilistic model for cell tower evaluation that would not require field testing for the evaluation of cell tower connections and have presented first results in (Bosma, 2022). The author of the present work is somewhat sceptical whether complete independence from in situ measurements will be possible, as the observed behaviour within this thesis is chaotic to a degree that surpasses simple description. However, if the model could be validated, this would greatly reduce the work required to evaluate cell tower connections. Failing a general validation, approaches could be developed allowing for the validation of the

model to a specific situation using a predefined protocol, or the model could be built in a way that specific measures would allow a calibration for a given situation.

Finally, the measurements conducted in this work were done completely manually, by restarting, calling and noting the current cell tower by hand. For further research, it should be considered to look into the automation of these measurement processes.

8.3.3 A-GPS / Location Services

Similarly as with cell towers, there is little empiric research into what impacts the result of the location services for Apple and Google. Again, it is difficult to know what consists comparable conditions for reference data creation. In addition to the factors to be studied for the cell towers, the effect of changing networks should be investigated for location services. As A-GPS is partially based on databases of cell towers and WiFi-access points, the behaviour of these services when cell towers or access points appear or disappear should be investigated.

The model for evaluation presented in Chapter 6 seems promising to be adaptable for general use. Further improvements could be achieved by modeling the probability as a two-dimensional distribution instead of splitting up the probabilities for angle and distance. A larger scale validation and calibration study should be conducted to test the model. As with the Cell Towers, the possibility to model the location without on site measurements should be investigated, however due to the even higher complexity of the process, it is even less likely this will be possible.

The possibility to automate the reference data collection should be considered.

8.3.4 Behavioural Biometrics

Whilst showing potential for one to one comparison, the here presented behavioural biometrics approach still has quite a way to go for comparison against a larger population. Future research should focus on improving the efficiency and the discriminatory power of the method by studying alternative characteristics. Research should be conducted with a large scale data set containing data from different ages, professions, genders and socio-economic background to study the influence of these factors on device usage.

So far, the approach only gives useful information if a high degree of correspondence is observed, as there is a plethora of reasons why such a difference may come to be. A change in daily activity, such as vacation, a

new job or an especially stressful day at work may cause activity to change. Also, behaviour may change with varying mental well being[3]. Additionally, the question persists whether the approach also works for period during which criminal activity takes place. Whilst for serial perpetrators, committing a crime may consist «normal» behaviour, the majority of criminals are single time offenders (Kuhn, 2012). As such, criminal behaviour, per definition, is extraordinary behaviour. It is currently unknown whether this is also reflected in phone usage. A study specifically addressing this aspect would be required.

Beyond identifying a person, there may be other uses for behavioural biometric analysis. As the method is signaling different behaviour, the approach may also be used to find periods of anomalous activity and identify time frames of interest such as deviations from a person's routine. Showing consistent behaviour over a given time frame may be used to authenticate digital traces as it is unlikely that a altered data set would manage to be consistent. Further research would help developing approaches adapted for these situations.

8.3.5 Likelihood Ratio

In this work, some general issues related to LR and probabilities have come up. Especially for propositions that are not actually true, trace values are likely to be quite a bit outside of the reference values. Whilst this is in principle a good thing, as it indicates that the observed characteristic is well adapted to distinguish between the propositions at hand, it poses a problem regarding the stability of the probability of the evidence under this proposition. Indeed, small variations in the data may cause a change of order of magnitude in the probability. Additionally, probabilities of 10^{-6} or lower are very hard to defend if they are based on just about a hundred of measurements. Further research should be conducted to develop approaches to handle such situations.

One approach, used in this work is defining lower bound probabilities below which the expert does not feel justified to go. Although not encountered in this thesis, the same issue may arise with very high values. For example, if the device-level LR would be approaching infinity, the overall LR is likely to become unstable. It would be useful to have guidelines supporting experts in choosing these lower or upper bound values.

[3]Psychotherapists have proposed methods of predicting such phases based on changed behaviour for therapeutic uses (see for example (Messner et al., 2019).

8.4 Conclusion

Location-related mobile device evidence is increasingly being used to address forensic questions in criminal investigations, particularly the location of a person alleged to have possessed and / or operated the device during the time of interest. A significant part of the forensic science community considers the distinction between device and person to be essential. However, some practitioners incorrectly present person-location information as fact rather than interpretation. Others neglect their duty to render an expert opinion by simply reporting device-level information, leaving it to non-specialists (e.g., prosecutor, investigator, jury) to form their own non-scientific opinion on the person-device link. Some practitioners attempt to establish general ownership of the device, but do not use a robust, repeatable, method. More importantly, practitioners do not have a structured framework for formulating and expressing scientific evaluation of mobile device evidence under person-level, location-focused propositions. It is necessary to properly structure and interpret this evidence to avoid mistakes, misinterpretations, and miscarriages of justice.

This work makes a large step into the relatively young field of the interpretation of location-related digital evidence. This work provides a robust framework that clearly distinguishes between "what has been observed" (i.e., what data are available) and how those data may inform about uncertain propositions in the case at hand. Specifically, this work structures the problem and provides a Bayesian framework to handle uncertainties in the context of the problem. The proposed approach also provides a possibility to address the person-device gap that may be applicable to other situations where this issue arises. In four simulated case scenarios, it was shown that it is possible to apply the approach to situations as one may encounter in real world scenarios. This work also provides guidance for practitioners to apply the framework to actual cases, enabling them to evaluate location-related evidence and uncertainties resulting from the Person-Device gap in a logical consistent manner, supporting decision-makers with balanced and founded results.

Overall, this work provides the following contributions to existing research:

- The problem is analysed in detail and structured.
- Means to close the Person-Device Gap are discussed.
- A Bayesian approach to address uncertainties resulting from the Person-Device Gap is presented.

- A Bayesian network for the evaluation of location-related evidence on person-level is proposed, studied and tested.
- The possibility to distinguish between two users on the same device is demonstrated.
- It is shown that it is possible to quantify uncertainties linked to digital traces.
- A rudimentary approach for the evaluation of cell tower evidence is proposed.
- An approach for the evaluation of device localisations is presented.
- Issues related to the application of this approach in real world cases are discussed.

Large parts of this work being exploratory, many questions and issues encountered remain unresolved. Especially research related to the workings of location services, the viability of behavioural biometrics in real world cases and the integration of uncertainties of the preliminary stage remain a challenge. However, this work provides a solid foundation to address these questions and motivation for future research.

Bibliography

Colin Aitken, Anders Nordgaard, Franco Taroni, and Alex Biedermann. Commentary: Likelihood Ratio as Weight of Forensic Evidence: A Closer Look. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00224. URL https://www.frontiersin.org/articles/10.3389/fgene.2018.00224/full. Publisher: Frontiers.

Eesa Al Solami, Colin Boyd, Andrew Clark, and Asadul K. Islam. Continuous Biometric Authentication: Can It Be More Practical? In *2010 IEEE 12th International Conference on High Performance Computing and Communications (HPCC)*, pages 647–652, September 2010. doi: 10.1109/HPCC.2010.65.

Anja Evelyn Amundsen and Kenneth M. Ovens. Forensics analysis of Wi-Fi communication traces in mobile devices. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3632–3637, Boston, MA, USA, December 2017. IEEE. ISBN 978-1-5386-2715-0. doi: 10.1109/BigData.2017.8258357. URL http://ieeexplore.ieee.org/document/8258357/.

AOSP. Supporting Multiple Users, 2020. URL https://source.android.com/devices/tech/admin/multi-user.

Humaira Arshad, Aman Jantan, and Oludare Abiodun. Digital Forensics: Review of Issues in Scientific Validation of Digital Evidence. *Journal of Information Processing Systems*, 14:346 ~ 376, April 2018. doi: 10.3745/JIPS.03.0095.

Simon Baechler, Marie Morelato, Simone Gittelson, Simon Walsh, Pierre Margot, Claude Roux, and Olivier Ribaux. Breaking the barriers between intelligence, investigation and evaluation: A continuous approach to define the contribution and scope of forensic science. *Forensic Science International*, 309:110213, April 2020. ISSN 0379-0738. doi: 10.1016/j.forsciint.2020.110213. URL http://www.sciencedirect.com/science/article/pii/S037907382030075X.

Luisa Bassi and Aurèle Scoundrianos. Bayesian evaluation of digital location evidence: case report of a homicide investigation. In *EAFS Conference 2022*, Stockholm, May 2022.

Rev. Thomas Bayes and Richard Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763. ISSN 0260-7085. URL https://www.jstor.org/stable/105741. Publisher: The Royal Society.

Connie Bell. Providing Context to the Clues: Recovery and Reliability of Location Data from Android Devices. Master's thesis, University of Central Florida, Orlando, 2015. URL https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=2353&context=etd.

Bellingcat Investigation Team. Gun Safety, Self Defense, and Road Marches - Finding an ISIS Training Camp, August 2014. URL https://www.bellingcat.com/resources/case-studies/2014/08/22/gun-safety-self-defense-and-road-marches-finding-an-isis-training-camp/.

Bellingcat Investigation Team. How a Werfalli Execution Site Was Geolocated, October 2017. URL https://www.bellingcat.com/news/mena/2017/10/03/how-an-execution-site-was-geolocated/.

Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal*, 38 (3):218–226, September 2015. ISSN 1559-3126. doi: 10.1037/prj0000130. URL https://europepmc.org/articles/PMC4564327.

John Berry and David Stoney. History and Development of Fingerprinting. In Henry C. Lee, Robert Ramotowski, and R. E. Gaensslen, editors, *Advances in Fingerprint Technology*. CRC Press, Boca Raton, Fla, 2nd edition edition, June 2001.

Chandrasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywe, Lorrie Cranor, and Marios Savvides. *Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption.* January 2015. doi: 10.14722/usec.2015.23003.

Shanthi Bhatt and T. Santhanam. Keystroke dynamics for biometric authentication — A survey. In *2013 International Conference on Pattern*

Recognition, Informatics and Mobile Engineering, pages 17–23, February 2013. doi: 10.1109/ICPRIME.2013.6496441.

Alex Biedermann and Kyriakos N. Kotsoglou. Digital evidence exceptionalism? A review and discussion of conceptual hurdles in digital evidence transformation. *Forensic Science International: Synergy*, 2:262–274, January 2020. ISSN 2589-871X. doi: https://doi.org/10.1016/j.fsisyn.2020.08.004.

Alex Biedermann and Joëlle Vuille. Digital evidence, 'absence' of data and ambiguous patterns of reasoning. *Digital Investigation*, 16:S86–S95, 2016. ISSN 1742-2876.

Geoffrey Blewitt. Basics of the GPS Technique: Observation Equations, 1997.

Timothy Bollé, Eoghan Casey, and Maëlig Jacquet. The role of evaluations in reaching decisions using automated systems supporting forensic analysis. *Forensic Science International: Digital Investigation*, 34: 301016, September 2020a. ISSN 2666-2817. doi: 10.1016/j.fsidi.2020.301016. URL `http://www.sciencedirect.com/science/article/pii/S2666281720300755`.

Timothy Bollé, Francesco Servida, Johann Polewczyk, Thomas Souvignet, and Eoghan Casey. Expressing evaluative conclusions in cases involving tampering of digital evidence, June 2020b. URL `https://dfrws.org/wp-content/uploads/2020/06/DFRWS-EU-2020-Expressing-evaluative-conclusions-in-cases-involving-tampering-c pdf`.

Wauter Bosma. A probabilistic approach to estimating cell service areas. In *EAFS Conference 2022*, Stockholm, May 2022.

Wauter Bosma, Sander Dalm, Erwin van Eijk, Rachid el Harchaoui, Edwin Rijgersberg, Hannah Tereza Tops, Alle Veenstra, and Rolf Ypma. Establishing phone-pair co-usage by comparing mobility patterns. *Science & Justice*, 60(2):180–190, March 2020. ISSN 1355-0306. doi: 10.1016/j.scijus.2019.10.005. URL `https://www.sciencedirect.com/science/article/pii/S1355030619300942`.

Alan S. Brown, Elisabeth Bracken, Sandy Zoccoli, and King Douglas. Generating and remembering passwords. *Applied Cognitive Psychology*, 18(6):641–651, 2004. ISSN 1099-0720. doi: 10.1002/acp.1014.

URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.1014.

Ann W. Bunch. National Academy of Sciences "Standardization": On What Terms? *Journal of Forensic Sciences*, 59(4):1041–1045, July 2014. ISSN 00221198. doi: 10.1111/1556-4029.12496. URL http://doi.wiley.com/10.1111/1556-4029.12496.

Matt Burgess. To protect Putin, Russia is spoofing GPS signals on a massive scale. *Wired UK*, March 2019. ISSN 1357-0978. URL https://www.wired.co.uk/article/russia-gps-spoofing. Section: Russia.

Mark Burnett. 10 million Passwords, February 2015a. URL http://archive.org/details/10MillionPasswords.

Mark Burnett. Ten Million Passwords FAQ, February 2015b. URL https://xato.net/ten-million-passwords-faq-3b2752ed3b4c.

Mark Burnett. Today I Am Releasing Ten Million Passwords, February 2015c. URL https://xato.net/today-i-am-releasing-ten-million-passwords-b6278bbe7495.

Eoghan Casey. Error, Uncertainty and Loss in Digital Evidence. *IJDE*, 1, January 2002.

Eoghan Casey. Standardization of Forming and Expressing Preliminary Evaluative Opinions on Digital Evidence. *Digital Investigation*, 2020.

Eoghan Casey and David-Olivier Jaquet-Chiffelle. Do Identities Matter? *Policing: A Journal of Policy and Practice*, 13(1):21–34, June 2017. ISSN 1752-4512. doi: 10.1093/police/pax034. URL https://doi.org/10.1093/police/pax034.

Eoghan Casey and Benjamin Turnbull. Digital Evidence on Mobile Devices. In *Digital Evidence and Computer Crime*, page 44. Academic Press, Cambridge, Massachusetts, 3rd edition edition, 2011.

Eoghan Casey, David-Olivier Jaquet-Chiffelle, Hannes Spichiger, Elénore Ryser, and Thomas Souvignet. Structuring the Evaluation of Location-Related Mobile Device Evidence. *Forensic Science International: Digital Investigation*, 32:300928, April 2020a. ISSN 2666-2817. doi: 10.1016/j.fsidi.2020.300928. URL https://www.sciencedirect.com/science/article/pii/S2666281720300238.

Eoghan Casey, Hannes Spichiger, Elénore Ryser, Francesco Servida, and David-Olivier Jaquet-Chiffelle. IoT Forensic Science: Principles, Processes, and Activities. In Parag Chatterjee, Emmanuel Benoist, and Asoke Nath, editors, *Applied Approach to Privacy and Security for the Internet of Things*, pages 1–37. IGI Global, Hershey, PA, 2020b. ISBN 978-1-79982-444-2. doi: 10.4018/978-1-7998-2444-2.ch001. URL `www.igi-global.com/chapter/iot-forensic-science/257902`.

Joakim Cedergren. Assisted GPS for Location Based Services. Master's thesis, Blekinge Institute of Technology, 2005. URL `http://www.diva-portal.org/smash/get/diva2:833206/FULLTEXT01`.

Christophe Champod and Massimo Tistarelli. Biometric Technologies for Forensic Science and Policing: State of the Art. In Massimo Tistarelli and Christophe Champod, editors, *Handbook of Biometrics for Forensic Science*, pages 1–15. Springer International Publishing, Cham, 2017. ISBN 978-3-319-50671-5 978-3-319-50673-9. doi: 10.1007/978-3-319-50673-9_1. URL `http://link.springer.com/10.1007/978-3-319-50673-9_1`.

Christophe Champod, Chris Lennard, Pierre Margot, and Milutin Stoilovic. *Traces et empreintes digitales: traité de dactyloscopie*. Presses polytechniques et universitaires romandes, 2017. ISBN 978-2-88915-183-7. Google-Books-ID: iHvAAQAACAAJ.

Circuit Court of Albermarle County. Commonwealth of Virginia v. M. L. Weiner, July 2015. Docket-Nr: CR000030-00.

CJEU. G.D. v Commissioner of An Garda Síochána, Minister for Communications, Energy and Natural Resources, Attorney General,, May 2022. URL `https://curia.europa.eu/juris/document/document.jsf;jsessionid=A4FA0D5F634BAAC37DB9A616FEBD8765?text=&docid=257242&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=235415`.

Roger Cook, Ian Evett, Graham Jackson, Phil Jones, and Jim Lambert. A hierarchy of propositions: Deciding which level to address in casework. *Science & Justice*, 38:231–239, October 1998. doi: 10.1016/S1355-0306(98)72117-3.

Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The Tangled Web of Password Reuse. In *Proceedings 2014 Network and Distributed System Security Symposium*, San Diego, CA, 2014. Internet Society. ISBN 978-1-891562-35-8. doi: 10.14722/ndss.2014.

23357. URL https://www.ndss-symposium.org/ndss2014/programme/tangled-web-password-reuse/.

Dipankar Dasgupta, Arunava Roy, and Abhijit Nag. Authentication Basics. In Dipankar Dasgupta, Arunava Roy, and Abhijit Nag, editors, *Advances in User Authentication*, pages 1–36. 2017. doi: 10.1007/978-3-319-58808-7_1. Publisher: Springer, Cham.

Kim De Bie. Assessing evidence for shared ownership of two phones using a statistical model and call detail records. In *EAFS Conference 2022*, Stockholm, May 2022.

Rachna Dhamija and Adrian Perrig. Déjà Vu: A User Study Using Images for Authentication. *USENIX Security Symposium*, 9, 2000.

Nicole M Egli. *Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System.* Doctorate Thesis, School of Criminal Justice, University of Lausanne, Lausanne, 2009.

Manuel Eichelberger, Ferdinand von Hagen, and Roger Wattenhofer. A Spoof-Proof GPS Receiver. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 145–156, April 2020. doi: 10.1109/IPSN48710.2020.00-39.

ENFSI. ENFSI Guideline for Evaluative Reporting in Forensic Science. Technical report, European Network of Forensic Science Institutes, 2010. URL http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

EuroPol. Stop Child Abuse – Trace an Object, June 2017. URL https://www.europol.europa.eu/stopchildabuse.

EWCA. Calland, R. v [2017] EWCA Crim 2308, December 2017. URL http://www.bailii.org/ew/cases/EWCA/Crim/2017/2308.html.

EWCA. Turner, R. v [2020] EWCA Crim 1241, September 2020. URL http://www.bailii.org/ew/cases/EWCA/Crim/2020/1241.html.

Ramsey Faragher and Robert Harle. An Analysis of the Accuracy of Bluetooth Low Energy for Indoor Positioning Applications. In *Proceedings of the 27th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2014)*, pages 201–210, September 2014. URL http://www.ion.org/publications/abstract.cfm?jp=p&articleID=12411. ISSN: 2331-5954.

FBI. 2017 Biometric Identification Award, February 2018. URL https://www.fbi.gov/video-repository/biometric_award_2017.mp4/view.

Michael O. Finkelstein and William B. Fairley. A Bayesian Approach to Identification Evidence. *Harvard Law Review*, 83(3):489–517, 1970. ISSN 0017-811X. doi: 10.2307/1339656. URL https://www.jstor.org/stable/1339656.

Giancarlo Fiorella. Geolocating Venezuelan Lawmakers In Europe, January 2020. URL https://www.bellingcat.com/news/2020/01/21/geolocating-venezuelan-lawmakers-in-europe/.

Dinei Florencio and Cormac Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, page 657, Banff, Alberta, Canada, 2007. ACM Press. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242661. URL http://portal.acm.org/citation.cfm?doid=1242572.1242661.

Federal Office of Public Health FOPH. Coronavirus: SwissCovid app and contact tracing, October 2020. URL https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/swisscovid-app-und-contact-tracing.html.

Forensic Focus. iPhone Tracking - from a forensic point of view, November 2011. URL https://www.forensicfocus.com/articles/iphone-tracking-from-a-forensic-point-of-view/.

FSR. Codes of Practice and Conduct, Appendix: Digital Forensics – Cell Site Analysis. Technical Report FSR-C-135, UK Forensic Science Regulator, Birmingham, 2020. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/918946/135_FSR-C-135_Cell_Site_Analysis_Issue_2.pdf.

FSR. Development of evaluative opinions. Technical Report FSR-C-118, UK Forensic Science Regulator, Birmingham, 2021. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/960051/FSR-C-118_Interpretation_Appendix_Issue_1__002_.pdf.

Xinwen Fu, Nan Zhang, Aniket Pingley, Wei Yu, Jie Wang, and Wei Zhao. The Digital Marauder's Map: A WiFi Forensic Positioning Tool. *IEEE Transactions on Mobile Computing*, 11(3):377–389, March 2012. ISSN

1536-1233. doi: 10.1109/TMC.2011.70. URL http://ieeexplore.ieee.org/document/5740913/.

Julien Furrer, Romain Voisard, Christophe Champod, and Marco De Donno. PiAnoS documentation — PiAnoS - Picture Annotation System, 2020. URL https://ips-labs.unil.ch/doc/.

Christopher Galbraith, Padhraic Smyth, and Hal Stern. Statistical Methods for the Forensic Analysis of Geolocated Event Data. *Forensic Science International: Digital Investigation*, 33:301009, July 2020. doi: 10.1016/j.fsidi.2020.301009.

Robert Gavin. Judge blocks Google evidence from Troy murder trial, October 2017. URL https://www.timesunion.com/news/article/Google-evidence-tossed-from-Troy-suitcase-murder-12311986.php. Section: News.

Shirley Gaw and Edward W Felten. Password Management Strategies for Online Accounts. In *Proceedings of the 2nd Symposium on Usable Privacy and Security*, page 12, Pitttsburgh, Pennsylvania, 2006. doi: 10.1145/1143120.1143127. URL https://cups.cs.cmu.edu/soups/2006/proceedings/p44_gaw.pdf.

Google and Harris Poll. Online Security Survey. Technical report, Google, February 2019. URL https://services.google.com/fh/files/blogs/google_security_infographic.pdf.

Martin Griffiths and Joe Hoy. RFPS: Scanners vs Test Phones. Technical Briefing 01674-BRF, Forensic Analytics Ltd, Letchworth, September 2018. URL https://www.forensicanalytics.co.uk/wp-content/uploads/2018/09/0164-BRF-RFPS-scanners-vs-test-phones-v2.0.pdf.

Mark Guido, Marc Brooks, Justin Grover, Eric Katz, Jared Ondricek, Marcus Rogers, and Lauren Sharpe. Generating a Corpus of Mobile Forensic Images for Masquerading user Experimentation. *Journal of Forensic Sciences*, 61(6):1467–1472, 2016. ISSN 1556-4029. doi: 10.1111/1556-4029.13178. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.13178. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1556-4029.13178.

Mo Harber-Lamond. How to spoof your location for Pokémon GO on Android, October 2020. URL https://www.tomsguide.com/how-to/how-to-spoof-your-location-for-pokemon-go-on-android.

Eliot Higgins. A Beginner's Guide to Geolocating Videos, July 2014. URL https://www.bellingcat.com/resources/how-tos/2014/07/09/a-beginners-guide-to-geolocation/.

Graeme Horsman. Digital Evidence Certainty Descriptors (DECDs). *Forensic Science International: Digital Investigation*, 32:200896, March 2020. ISSN 2666-2817. doi: https://doi.org/10.1016/j.fsidi.2019.200896.

Wiger Houten, Ivo Alberink, and Zeno Geradts. Implementation of the Likelihood Ratio framework for camera identification based on sensor noise patterns. *Law, Probability and Risk*, 10:149–159, June 2011. doi: 10.1093/lpr/mgr006.

Joseph Hoy. Forensic Radio Surveys for Cell Site Analysis. In *Forensic Radio Survey Techniques for Cell Site Analysis*, pages 1–2. Wiley, 2015. ISBN 978-1-118-92574-4. doi: 10.1002/9781118925768.ch1. URL https://ieeexplore.ieee.org/document/8043829.

Roy A. Huber. Expert Witnesses. *Criminal Law Quarterly*, 3:276–295, 1959. URL https://heinonline.org/HOL/Page?handle=hein.journals/clwqrty2&id=463&div=&collection=.

Hugin Expert. Hugin HDE. URL https://www.hugin.com.

Sohaib Ikram and Hafiz Malik. Digital audio forensics using background noise. In *2010 IEEE International Conference on Multimedia and Expo*, pages 106–110, Singapore, Singapore, July 2010. IEEE. ISBN 978-1-4244-7491-2. doi: 10.1109/ICME.2010.5582981. URL http://ieeexplore.ieee.org/document/5582981/.

ISDC. Gutachten über Strafprozessuale Grundlagen und Kosten der Überwachung von Fernmeldeverkehr in Dänemark, Deutschland, Frankreich, Italien, den Niederlanden und dem Vereinigten Königreich. Technical Report Avis 12-051, Institut Suisse de Droit Comparé, Lausanne, May 2013. URL https://www.li.admin.ch/documents/site/SIR_Gutachten_20130524.pdf.

ISO/TC 272 Forensic sciences. ISO/CD 21043-4 Forensic Sciences — Part 4: Interpretation. Technical Report 1, ISO. URL https://www.iso.org/standard/72039.html?browse=tc.

Blake Ives, Kenneth R. Walsh, and Helmut Schneider. The domino effect of password reuse. *Communications of the ACM*, 47(4):75–78, April 2004.

ISSN 0001-0782, 1557-7317. doi: 10.1145/975817.975820. URL https://dl.acm.org/doi/10.1145/975817.975820.

David-Olivier Jacquet-Chiffelle. The Model : A Formal Description, Section 7.2. In David-Olivier Jacquet-Chiffelle, Bernhard Anrig, Emmanuel Benoist, and Rolf Haenni, editors, *Virtual Persons and Identities*, FIDIS deliverable 2.13. September 2008. URL http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp17-del17.1.Modelling_New_Forms_of_Identities.pdf.

Emad Sami Jaha and Mark S. Nixon. From Clothing to Identity: Manual and Automatic Soft Biometrics. *IEEE Transactions on Information Forensics and Security*, 11(10):2377–2390, October 2016. ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2016.2584001. URL http://ieeexplore.ieee.org/document/7498567/.

Vladan M. Jovanovic and Brian T. Cummings. Analysis of Mobile Phone Geolocation Methods Used in US Courts. *IEEE Access*, 10:28037–28052, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3156892.

Aikaterini Kanta, Iwen Coisel, and Mark Scanlon. PCWQ: A Framework for Evaluating Password Cracking Wordlist Quality. *The 12th EAI International Conference on Digital Forensics and Cyber Crime*, December 2021. URL https://forensicsandsecurity.com/papers/PasswordCrackingWordlistQuality.php. Publisher: Springer.

Aikaterini Kanta, Iwen Coisel, and Mark Scanlon. A Novel Dictionary Generation Methodology for Contextual-Based Password Cracking. *IEEE Access*, 10:59178–59188, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3179701.

David D. Kirkpatrick. Who Is Behind QAnon? Linguistic Detectives Find Fingerprints. *The New York Times*, February 2022. ISSN 0362-4331. URL https://www.nytimes.com/2022/02/19/technology/qanon-messages-authors.html.

Kenneth Knowlson. Targeted advertising, May 2003. URL https://patents.google.com/patent/US20030093311A1/en.

Kjell Konis and Janhavi Moharil. RHugin, July 2008. URL https://rhugin.r-forge.r-project.org/.

Tomislav Kos, Ivan Markezic, and Josip Pokrajcic. Effects of multipath reception on GPS positioning performance. In *Proceedings ELMAR-2010*, pages 399–402, Zadar, Croatia, 2010. IEEE. ISBN 978-1-4244-6371-8.

Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. SpotFi: Decimeter Level Localization Using WiFi. *ACM SIGCOMM Computer Communication Review*, 45(4):269–282, August 2015. ISSN 0146-4833. doi: 10.1145/2829988.2787487. URL https://doi.org/10.1145/2829988.2787487.

André Kuhn. *Quel avenir pour la justice pénale?*, volume 76 of *la question*. Ed. de l'Hèbe, Grolley, 2012. ISBN 978-2-88906-046-7. Google-Books-ID: s56aMwEACAAJ.

Philipp Kuhn. Mordprozess Hussein K.: „Die Version vom Handeln im Affekt ist mit dem heutigen Tag obsolet". *DIE WELT*, January 2018. URL https://www.welt.de/vermischtes/article172287105/Mordprozess-Hussein-K-Die-Version-vom-Handeln-im-Affekt-ist-mit-dem-heutigen-T html.

Nicolai Kuntze, Carsten Rudolph, Aaron Alva, Barbara Endicott-Popovsky, John Christiansen, and Thomas Kemmerich. On the Creation of Reliable Digital Evidence. In *Advances in Digital Forensics VIII*, pages 3–17. Springer, Berlin, Heidelberg, January 2012. doi: 10.1007/978-3-642-33962-2_1. URL https://bib-ezproxy.epfl.ch:5050/chapter/10.1007/978-3-642-33962-2_1.

Michael Kwan, Kam-Pui Chow, Frank Law, and Pierre Lai. Reasoning About Evidence Using Bayesian Networks. In Indrajit Ray and Sujeet Shenoi, editors, *Advances in Digital Forensics IV*, pages 275–289, Boston, MA, USA, 2008. Springer US. ISBN 978-0-387-84927-0.

LastPass. The 3rd Annual Global Password Security Report. Technical Report 3, LastPass, 2019. URL https://www.lastpass.com/state-of-the-password/global-password-security-report-2019.

Shawn Levy. Chapter Three: The Case of the Missing Lifeguard, July 2019. IMDb ID: tt7911866 event-location: USA.

LexisNexis. LexisNexis Behavioral Biometrics, 2020. URL https://risk.lexisnexis.com/products/behavioral-biometrics.

Fudong Li, Nathan Clarke, Maria Papadaki, and Paul Haskell-Dowland. Behaviour Profiling for Transparent Authentication for Mobile Devices.

Tallinn, July 2011. Academic Publishing Ltd, UK. URL https://ro.ecu.edu.au/ecuworks2011/706/.

Leslie Libman. NCIS: Deception, January 2006. IMDb ID: tt0657990.

Edmond Locard. *L'enquête criminelle et les méthodes scientifiques*. Bibliothèque de philosophie scientifique. E. Flammarion, Paris, 1920. URL https://books.google.ch/books?id=fkUuAAAAYAAJ.

Lock Screen Master. Gesture Lock Screen – Apps on Google Play, September 2022. URL https://play.google.com/store/apps/details?id=qlocker.draw&hl=en_IN&gl=US.

Léon Lopez. Evaluation des recherches par champ d'antennes : Comparaison de différents instruments de mesure. Master's thesis, Ecole des Sciences Criminelles, University of Lausanne, Lausanne, June 2021.

Halgurd S. Maghdid, Ihsan Alshahib Lami, Kayhan Zrar Ghafoor, and Jaime Lloret. Seamless Outdoors-Indoors Localization Solutions on Smartphones: Implementation and Challenges. *ACM Computing Surveys*, 48(4):53:1–53:34, February 2016. ISSN 0360-0300. doi: 10.1145/2871166. URL https://doi.org/10.1145/2871166.

Hafiz Malik. Acoustic Environment Identification and Its Applications to Audio Forensics. *IEEE Transactions on Information Forensics and Security*, 8(11):1827–1837, November 2013. ISSN 1556-6021. doi: 10.1109/TIFS.2013.2280888.

Raymond Marquis, Alex Biedermann, Liv Cadola, Christophe Champod, Line Gueissaz, Geneviève Massonnet, Williams David Mazzella, Franco Taroni, and Tacha Hicks. Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, 56(5):364–370, September 2016. ISSN 1355-0306. doi: 10.1016/j.scijus.2016.05.009. URL https://www.sciencedirect.com/science/article/pii/S1355030616300338.

Krista Merry and Pete Bettinger. Smartphone GPS accuracy study in an urban environment. *PLoS ONE*, 14:e0219890, July 2019. doi: 10.1371/journal.pone.0219890.

Eva-Maria Messner, Rayna Sariyska, Benjamin Mayer, Christian Montag, Christopher Kannen, Andreas Schwerdtfeger, and Harald Baumeister. Future Implications of Passive Smartphone Sensing in the Therapeutic Context. *Verhaltenstherapie*, pages 1–10, August 2019. ISSN 1016-6262, 1423-

0402. doi: 10.1159/000501951. URL https://www.karger.com/Article/FullText/501951.

Gaëtan Michelet. Détecter un changement d'utilisateur sur un smartphone. Master's thesis, Ecole des Sciences Criminelles: University of Lausanne, Lausanne, 2021.

Alastair H. Moore, Mike Brookes, and Patrick A. Naylor. Roomprints for forensic audio applications. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, New Paltz, NY, USA, October 2013. IEEE. ISBN 978-1-4799-0972-8. doi: 10.1109/WASPAA.2013.6701854. URL http://ieeexplore.ieee.org/document/6701854/.

Cedric Neumann, Ismael Mateos-Garcia, Glenn Langenburg, Jennifer Kostroski, James E. Skerrett, and Martin Koolen. Operational benefits and challenges of the use of fingerprint statistical models: A field study. *Forensic Science International*, 212(1):32–46, October 2011. ISSN 0379-0738. doi: 10.1016/j.forsciint.2011.05.004. URL https://www.sciencedirect.com/science/article/pii/S0379073811002192.

OpenStreetMap. OpenStreetMap. URL https://www.openstreetmap.org/.

Richard E. Overill and Jan Collie. Quantitative evaluation of the results of digital forensic investigations: a review of progress. *Forensic Sciences Research*, pages 1–6, February 2021. ISSN 2096-1790, 2471-1411. doi: 10.1080/20961790.2020.1837429. URL https://www.tandfonline.com/doi/full/10.1080/20961790.2020.1837429.

Richard E. Overill and Jantje A. M. Silomon. Uncertainty Bounds for Digital Forensic Evidence and Hypotheses. pages 590–595. IEEE Computer Society, August 2012. ISBN 978-0-7695-4775-6. doi: 10.1109/ARES.2012.17. URL https://dl.acm.org/doi/10.1109/ARES.2012.17.

Richard E. Overill, Jantje A. M. Silomon, Michael Y. K. Kwan, Kam-Pui Chow, Frank Y. W. Law, and Pierre K. Y. Lai. Sensitivity Analysis of a Bayesian Network for Reasoning about Digital Forensic Evidence. pages 1–5, Cebu, August 2010. IEEE. doi: 10.1109/HUMANCOM.2010.5563318. URL https://ieeexplore.ieee.org/document/5563318.

Richard E. Overill, Jantje A. M. Silomon, Kam-Pui Chow, and Rayson Tse. Quantification of digital forensic hypotheses using probability theory. pages 1–5, Hong Kong, November 2013. IEEE. doi: 10.1109/SADFE.2013.6911547. URL https://ieeexplore.ieee.org/document/6911547.

Rashmi Patil, Rashmika K. Patole, and Priti P. Rege. Audio Environment Identification. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5, Kanpur, India, July 2019. IEEE. ISBN 978-1-5386-5906-9. doi: 10.1109/ICCCNT45670.2019.8944427. URL https://ieeexplore.ieee.org/document/8944427/.

Charles S. Pierce. Illustrations of the Logic of Science IV. In *Popular Science Monthly Volume 12 April 1878*. New York: D. Appleton & Co., April 1877.

PlaceIQ. Location Data Accuracy Revealed, August 2016. URL https://www.placeiq.com/wp-content/uploads/2019/11/PlaceIQ-Location-Data-Accuracy-Study_Findyr-White-Paper.pdf.

Robert Pless, A. Stylianou, and Austin Abrams. *Finding Jane Doe: A forensic application of 2D image calibration*. 2013. ISBN 978-1-84919-904-9.

Mark Pollitt, Eoghan Casey, David-Olivier Jaquet-Chiffelle, and Pavel Gladyshev. A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence. Technical report, OSAC, https://www.nist.gov/system/files/documents/2018/01/10/osac_ts_0002.pdf, 2018.

Nicole Poser. Judge blocks Google evidence from Troy murder trial, November 2017. URL https://www.hawkanalytics.com/judge-blocks-google-evidence-from-troy-murder-trial/.

PTSS. Post and Telecommunications Surveillance Service | Post and Telecommunications Surveillance Service PTSS, September 2021. URL https://www.li.admin.ch/en/.

Q Locker. Gesture Lock Screen - Apps on Google Play, March 2022. URL https://play.google.com/store/apps/details?id=qlocker.gesture&hl=en&gl=CH.

O. Ribaux. *Police scientifique: Le renseignement par la trace*. Sciences forensiques. PPUR, Presses polytechniques et universitaires romandes, 2014. ISBN 978-2-88915-061-8. URL https://books.google.ch/books?id=03DhoAEACAAJ.

Shannon Riley. Password Security: What Users Know and What They Actually Do. *Usability News*, 8(1):5, 2006.

Andrea Macarulla Rodriguez, Christian Tiberius, Roel van Bree, and Zeno Geradts. Google timeline accuracy assessment and error prediction. *Forensic Sciences Research*, 3(3):240–255, July 2018. ISSN 2096-1790. doi: 10.1080/20961790.2018.1509187. URL https://doi.org/10.1080/20961790.2018.1509187. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/20961790.2018.1509187.

Michael Leonard Rogers. Multi-Factor Authentication Using a Mobile Phone, June 2011. URL https://patents.google.com/patent/US20110142234A1/en.

Elénore Ryser and David-Olivier Jacquet-Chiffelle. Accuracy of geolocation metadata on pictures taken using a mobile phone, December 2021. URL https://dfrws.org/presentation/accuracy-of-geolocation-metadata-on-pictures-taken-using-a-mobile-phone/.

Elénore Ryser, Hannes Spichiger, and Eoghan Casey. Structured decision making in investigations involving digital and multimedia evidence. *Forensic Science International: Digital Investigation*, 34:301015, September 2020. ISSN 2666-2817. doi: 10.1016/j.fsidi.2020.301015. URL http://www.sciencedirect.com/science/article/pii/S2666281720300512.

Hataichanok Saevanee, Nathan Clarke, Steven Furnell, and Valerio Biscione. Continuous user authentication using multi-modal biometrics. *Computers & Security*, 53:234–246, September 2015. ISSN 0167-4048. doi: 10.1016/j.cose.2015.06.001. URL https://www.sciencedirect.com/science/article/pii/S0167404815000875.

Tuomas Salonen. Measuring the understandability of forensic reports: a questionnaire based test, February 2022.

Brittany Shammas. A man walked down a street with 99 phones in a wagon. Google Maps thought it was a traffic jam. *Washington Post*, February 2020. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2020/02/04/google-maps-simon-weckert/.

Hannes Spichiger. Interprétation des attributions de caméra source. Master's thesis, University of Lausanne, Lausanne, 2017.

Hannes Spichiger. The Use of Object Traces in a Connected World. Sydney, January 2021.

SPTA. Federal Act on the Surveillance of Post and Telecommunications (SPTA), March 2016. URL https://www.fedlex.admin.ch/eli/cc/2018/31/en.

Tim St. Onge. The Geographical Oddity of Null Island | Worlds Revealed: Geography & Maps at The Library Of Congress, April 2016. URL //blogs.loc.gov/maps/2016/04/the-geographical-oddity-of-null-island/.

Nina Sunde and Itiel E. Dror. Cognitive and human factors in digital forensics: Problems, challenges, and the way forward. *Digital Investigation*, 29:101–108, June 2019. ISSN 1742-2876. doi: 10.1016/j.diin.2019.03.011. URL http://www.sciencedirect.com/science/article/pii/S1742287619300441.

Superior Court of Pennsylvania. Commonwealth of Pennsylvania, Appellant v. Tyler Kristian Mangel, March 2018.

Swiss Federal Court. ATF 6B_1074/2018, January 2019. URL https://www.bger.ch/ext/eurospider/live/de/php/aza/http/index.php?highlight_docid=aza%3A%2F%2F24-01-2019-6B_1074-2018&lang=de&type=show_document&zoom=YES&.

Swiss Federal Criminal Court. BH.2014.16 + BP.2014.59, November 2014. URL https://bstger.weblaw.ch/pdf/20141106_BH_2014_16.pdf.

swisstopo. Swiss Geoportal. URL https://map.geo.admin.ch.

Zainab Syed, Jacques Georgy, Abdelrahman Ali, Hsiu-Wen Chang, and Chris Goodall. Showing Smartphones the Way Inside: Real-time, Continuous, Reliable, Indoor/Outdoor Localization. *GPS World*, 24:30–35, March 2013.

F. Taroni, A. Biedermann, S. Bozza, P. Garbolino, and C. Aitken. *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. Wiley, 2014. ISBN 978-0-470-97973-0. URL https://serval.unil.ch/notice/serval:BIB_0F30DFF15E4C.

Franco Taroni and Colin Aitken. Bayesian Networks and Probabilistic Inference in Forensic Science. May 2006. ISSN 9780470091739. doi: 10.1002/0470091754.ch4.

Matt Tart. Opinion evidence in cell site analysis. *Science & Justice*, 60 (4):363–374, July 2020. ISSN 1355-0306. doi: 10.1016/j.scijus.2020.02.002. URL http://www.sciencedirect.com/science/article/pii/S1355030619302898.

Matt Tart, Sue Pope, David Baldwin, and Robert Bird. Cell site analysis: Roles and interpretation. *Science & Justice*, 59(5):558–564, September 2019. ISSN 1355-0306. doi: 10.1016/j.scijus.2019.06.005. URL http://www.sciencedirect.com/science/article/pii/S1355030618303538.

Matt Tart, Iain Brodie, Nicholas Patrick-Gleed, Brian Edwards, Kevin Weeks, Robert Moore, and Richard Haseler. Cell site analysis; use and reliability of survey methods. *Forensic Science International: Digital Investigation*, 38:301222, September 2021. ISSN 2666-2817. doi: 10.1016/j.fsidi.2021.301222. URL https://www.sciencedirect.com/science/article/pii/S266628172100130X.

Matthew Tart, Iain Brodie, Nicholas Gleed, and James Matthews. Historic cell site analysis – Overview of principles and survey methodologies. *Digital Investigation*, 8(3):185–193, February 2012. ISSN 1742-2876. doi: 10.1016/j.diin.2011.10.002. URL http://www.sciencedirect.com/science/article/pii/S1742287611000867.

Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 85–98, New York, NY, USA, November 2009. Association for Computing Machinery. ISBN 978-1-60558-519-2. doi: 10.1145/1644038.1644048. URL https://doi.org/10.1145/1644038.1644048.

William C. Thompson, Ph D. Franco Taroni, Ph D, and Colin G. G. Aitken. How the probability of a false positive affects the value of DNA evidence. *Journal of Forensic Sciences*, pages 47–54, 2003.

William C. Thompson, Joelle Vuille, Alex Biedermann, and Franco Taroni. The role of prior probability in forensic assessments. *Frontiers in Genetics*, 4, 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00220. URL https://www.frontiersin.org/articles/10.3389/fgene.2013.00220/full. Publisher: Frontiers.

Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Buginon, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, Ludovic Barman, Sylvain Chatel, Kenneth Paterson, Srdjan Čapkun, David Basin, Jan Beutel, Dennis Jackson, Marc Roeschlin, Patrick Leu, Bart Preneel, Nigel Smart, Aysajan Abidin, Seda Gürses, Michael Veale, Cas Cremers, Michael Backes, Mils Ole

Tippenhauer, Reuben Binns, Ciro Cattuto, Alain Barrat, Dario Fiore, Manuel Barbosa, Rui Oliveira, and José Pereira. Decentralized Privacy-Preserving Proximity Tracing, May 2020. URL https://github.com/DP-3T/documents/blob/master/DP3T%20White%20Paper.pdf.

Hayson Tse, Kam-Pui Chow, and Michael Kwan. Reasoning about Evidence using Bayesian Networks. In Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics VIII*, IFIP Advances in Information and Communication Technology, pages 99–113, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-33962-2. doi: 10.1007/978-3-642-33962-2_7.

United States Department of Justice, and Federal Bureau of Investigation. *The Science of Fingerprints*. US Government Printing Office, Washington DC, 1984.

University of Lausanne. Planète UNIL. URL https://planete.unil.ch/.

Frank van Diggelen and Per Enge. The World's first GPS MOOC and Worldwide Laboratory using Smartphones. In *Proceedings of the 28th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2015)*, pages 361–369, Tampa, Florida, September 2015. URL https://www.ion.org/publications/abstract.cfm?articleID=13079.

Jan Van Sickle. Multipath, 2020. URL https://www.e-education.psu.edu/geog862/node/1721.

Chris Wood. WhatsApp photo drug dealer caught by 'groundbreaking' work. *BBC News*, April 2018. URL https://www.bbc.com/news/uk-wales-43711477.

Roman Yampolskiy and Venu Govindaraju. Behavioural biometrics: A survey and classification. *International Journal of Biometrics*, 1, January 2008. doi: 10.1504/IJBM.2008.018665.

Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. Digital multimedia audio forensics: past, present and future. *Multimedia Tools and Applications*, 77(1):1009–1040, January 2018. ISSN 1573-7721. doi: 10.1007/s11042-016-4277-2. URL https://doi.org/10.1007/s11042-016-4277-2.

H. Zhao and H. Malik. Audio Recording Location Identification Using Acoustic Environment Signature. *IEEE Transactions on Information Forensics and Security*, 8(11):1746–1759, November 2013. ISSN 1556-6021. doi:

10.1109/TIFS.2013.2278843. Conference Name: IEEE Transactions on Information Forensics and Security.

Hong Zhao and Hafiz Malik. Audio forensics using acoustic environment traces. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 373–376, Ann Arbor, MI, USA, August 2012. IEEE. ISBN 978-1-4673-0182-4 978-1-4673-0181-7. doi: 10.1109/SSP.2012.6319707. URL http://ieeexplore.ieee.org/document/6319707/.

Appendix A

Development of LR formula

This Appendix shows the development of the formula for the overall LR of the entire Bayes Net. This is not used in this work and provided for interested readers.

$$LR = \frac{Pr(E_1; E_2; E_3; E_4 \mid LocP_1; I)}{Pr(E_1; E_2; E_3; E_4 \mid LocP_2; I)} \tag{A.1}$$

The I representing the general information available to the expert in the case and will condition every probability throughout the entirety of the coming development. For the sake of brevity, it is omitted in the following and can be considered implied.

Using Bayes' theorem:

$$LR = \frac{Pr(E_1 \mid E_2; E_3; E_4; LocP_1)}{Pr(E_1 \mid E_2; E_3; E_4; LocP_2)} \cdot \frac{Pr(E_2; E_3; E_4 \mid LocP_1)}{Pr(E_2; E_3; E_4 \mid LocP_2)} \tag{A.2}$$

It can be shown, that in this particular situation, the probability of the states E_2, E_3 and E_4 are independent from the state of $LocP$. In other words

$$\begin{aligned} Pr(E_2; E_3; E_4 \mid LocP_1) &= Pr(E_2; E_3; E_4 \\ &\mid LocP_2) \\ &= Pr(E_2; E_3; E_4) \end{aligned} \tag{A.3}$$

As a consequence, the second fraction of formula A.2 reduces to one and the LR becomes as follows:

$$LR = \frac{Pr(E_1 \mid E_2; E_3; E_4; LocP_1)}{Pr(E_1 \mid E_2; E_3; E_4; LocP_2)} \tag{A.4}$$

A property of Bayesian Networks is that the probability of the states of a given node is only directly dependent from their parent nodes ("Screen

off" effect) (Taroni and Aitken, 2006). The screen off effect can be observed for any three nodes, A,B and C, in a divergent, convergent or consecutive relationship where node B is in any way linked to A and C whilst A and C are not directly dependent upon each other (such as shown in figure A.1). If in such a situation, the state of B is known, the state of A will no longer influence the probabilities of the state of C any longer and vice versa. Expressed as a formula, for any state A_n of node A and any state C_m of node C:

$$Pr(A_n \mid B, C) = Pr(A_n \mid B) \tag{A.5}$$

and

$$Pr(C_m \mid B, A_n) = Pr(C_m \mid B) \tag{A.6}$$

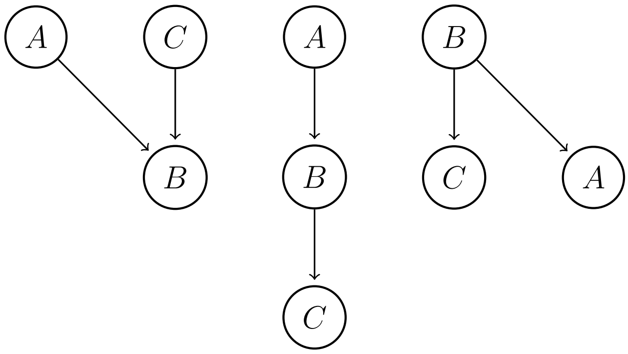


Figure A.1: Types of structures in a Bayesian network where screen off effects between A and C can be observed if the state of B is known.

For this particular situation, if the states of the *LocD*-Node are considered, the probabilities of E_1 can be expressed as only conditioned by the states of *LocD*. *LocD* has two states, $LocD_1$ (The device was located at location X at time t.) and $LocD_2$ (The device was located at location Y at time t.) which leads to the following development for the nominator:

$$Pr(E_1 \mid E_{2-4}; LocP_{1/2}) = \sum_{n=1}^{2} Pr(E_1 \mid LocD_n) \cdot Pr(LocD_n \mid E_{2-4}; LocP_{1/2}) \tag{A.7}$$

The probability for a given state of LocD is given as follows:

$$Pr(LocD_n \mid E_{2-4}; LocP_{1/2}) = \sum_{m=1}^{3} Pr(LocD_n \mid UseP_m; LocP_{1/2}; E_{2-4}) \cdot Pr(UseP_m \mid E_{2-4}) \quad (A.8)$$

The UseP-node is situated as such in the Bayesnet that a screen off effect exists between LocD and E_2 through E_4. Applying formula A.5, this dependency can be eliminated:

$$Pr(LocD_n \mid E_{2-4}; LocP_{1/2}) = \sum_{m=1}^{3} Pr(LocD_n \mid UseP_m; LocP_{1/2}) \cdot Pr(UseP_m \mid E_{2-4}) \quad (A.9)$$

The probability of any state of the node UseP given nodes E_2 through E_3 is obtained applying Bayes' theorem in its original form (cf. formula A.10).

$$Pr(A \mid B) = \frac{Pr(B \mid A) \cdot Pr(A)}{Pr(B)} \quad (A.10)$$

$$Pr(UseP_m \mid E_{2-4}) = \frac{Pr(E_2 \mid UseP_m; E_{3;4}) \cdot Pr(UseP_m \mid E_{3;4})}{Pr(E_2 \mid E_{3;4})} \quad (A.11)$$

The probability of E_2 can be expressed as the sum of it probabilities conditioned by the states of the UseP-node times the probability of those states:

$$Pr(UseP_m \mid E_{2-4}) = \frac{Pr(E_2 \mid UseP_m; E_{3;4}) \cdot Pr(UseP_m \mid E_{3;4})}{\sum_{m=1}^{3} Pr(E_2 \mid UseP_m; E_{3;4}) \cdot Pr(UseP_m \mid E_{3;4})} \quad (A.12)$$

With UseP being situated between the E_2-node and the E_3- and E_4-nodes, the probability, the dependency for E_2 from E_3 and E_4 can be eliminated:

$$Pr(UseP_m \mid E_{2-4}) = \frac{Pr(E_2 \mid UseP_m) \cdot Pr(UseP_m \mid E_{3;4})}{\sum_{m=1}^{3} Pr(E_2 \mid UseP_m) \cdot Pr(UseP_m \mid E_{3;4})} \quad (A.13)$$

The probability for each state of the UseP node can be expressed as follows:

$$Pr(UseP_m \mid E_{3;4}) = \frac{Pr(E_3 \mid UseP; E_4) \cdot Pr(UseP_m \mid E_4)}{Pr(E_3 \mid E_4)} \quad (A.14)$$

Where

$$Pr(UseP_m \mid E_4) = \sum_{l=1}^{3} Pr(UseP_m \mid User_l; E_4) \cdot Pr(User_l \mid E_4) \tag{A.15}$$

With the user-node being located between the UseP- and the E_4-node, the screen off effect allows to eliminate the direct dependencies between states of the two nodes:

$$Pr(UseP_m \mid E_4) = \sum_{l=1}^{3} Pr(UseP_m \mid User_l) \cdot Pr(User_l \mid E_4) \tag{A.16}$$

The probability of any state of the user-node given E_4 is as follows:

$$Pr(User_l \mid E_4) = \frac{Pr(E_4 \mid User_l) \cdot Pr(User_l)}{Pr(E_4)} \tag{A.17}$$

Which is equal to:

$$Pr(User_l \mid E_4) = \frac{Pr(E_4 \mid User_l) \cdot Pr(User_l)}{\sum_{l=1}^{3} Pr(E_4 \mid User_l) \cdot Pr(User_l)} \tag{A.18}$$

Additionnally:

$$Pr(E_3 \mid UseP_m; E_4) = \sum_{k=1}^{3} Pr(E_3 \mid UseU_k; UseP_m; E_4) \cdot Pr(UseU_k \mid UseP_m; E_4) \tag{A.19}$$

Using the screen-off-effect from the UseU-node situated between E_3 and other nodes, this simplifies to:

$$Pr(E_3 \mid UseP_m; E_4) = \sum_{k=1}^{3} Pr(E_3 \mid UseU_k) \cdot Pr(UseU_k \mid UseP_m; E_4) \tag{A.20}$$

Where:

$$Pr(UseU_k \mid UseP_m; E_4) = \sum_{l=1}^{3} Pr(useU_k \mid User_l; UseP_m; E_4) \cdot Pr(User_l \mid UseP_m; E_4) \tag{A.21}$$

Which (using the screen-off effect from the User-node situated between E_4 and the UseU-node) simplifies to:

$$Pr(UseU_k \mid UseP_m; E_4) = \sum_{l=1}^{3} Pr(useU_k \mid User_l; UseP_m) \cdot Pr(User_l \mid UseP_m; E_4) \tag{A.22}$$

Where:

$$Pr(User_l \mid UseP_m; E_3) = \frac{Pr(UseP_m \mid User_l; E_4) \cdot Pr(User_l \mid E_4)}{Pr(UseP_m \mid E_4)} \tag{A.23}$$

Also:

$$\begin{aligned} Pr(E_4 \mid E_3) &= \sum_{l=1}^{3} Pr(E_4 \mid E_3; User_l) \cdot Pr(User_l \mid E_3) \\ &= \sum_{l=1}^{3} Pr(E_4 \mid User_l) \cdot Pr(User_l \mid E_3) \end{aligned} \tag{A.24}$$

For the probability of the states of the User-node given E_3, the following development has to be made:

$$Pr(User_l \mid E_3) = \frac{Pr(E_3 \mid User_l) \cdot Pr(User_l)}{Pr(E_3)} \tag{A.25}$$

Where

$$Pr(E_3 \mid User_l) = \sum_{k=1}^{3} Pr(E_3 \mid UseU_k) \cdot Pr(UseU_k \mid User_l) \tag{A.26}$$

Where

$$Pr(UseU_k \mid User_l) = \sum_{m=1}^{3} Pr(UseU_k \mid UseP_m; User_l) \tag{A.27}$$

Also

$$Pr(E_3) = \sum_{k=1}^{3} Pr(E_3 \mid UseU_k) \cdot Pr(UseU_k) \tag{A.28}$$

Where

$$Pr(UseU_k) = \sum_{l=1}^{3} \sum_{m=1}^{3} Pr(UserU_k \mid User_l; UseP_m) \cdot Pr(User_l; UseP_m) \tag{A.29}$$

Which is equal to:

$$Pr(UseU_k) = \sum_{l=1}^{3} \sum_{m=1}^{3} Pr(UserU_k \mid User_l; UseP_m) \cdot Pr(UseP_m \mid User_l) \cdot Pr(User_l) \tag{A.30}$$

Inputting these formulae in the previous instances, a overall formula for the likelihood ratio can be obtained. for brevity reasons, this is not done here.

Appendix B

R-Scripts for Simulation of Bayes Nets

These R-scripts were used to simulate the behaviour of the Bayes Nets in Section 3.3. These scripts are available here: `https://github.com/HSpichig/Thesis/blob/main/BB_anonimize.py`. The Bayes Nets can be found here: `https://github.com/HSpichig/Thesis/blob/main/BN_Hugin.zip`

B.1 Influence of UseP

```
setwd("C:/Users/Hannes/switchdrive/Thesis")

library(RHugin)

BayesNet <- read.rhd("BN_Reduced_Sim1.net")

resolution = 100


LR <- function(BayesNet, useP1, useP2, e1_1, e1_2){

  tabUseP <- get.table(BayesNet, 'UseP')
  tabUseP[1,'Freq'] = useP1
  tabUseP[2,'Freq'] = useP2
  tabUseP[3,'Freq'] = 1-useP1-useP2
  set.table(BayesNet, 'UseP', tabUseP)

  tabE1 <- get.table(BayesNet, 'E1')
```

```
  tabE1[1,'Freq'] = e1_1
  tabE1[2,'Freq'] = 1-e1_1
  tabE1[3,'Freq'] = e1_2
  tabE1[4,'Freq'] = 1-e1_2
  set.table(BayesNet, 'E1', tabE1)

  compile(BayesNet)
  set.finding(BayesNet, 'E1', 'E1')
  propagate(BayesNet)
  odds <- get.belief(BayesNet, 'LocP')
  uncompile(BayesNet)

  lr <- odds['LocP1'] / odds['LocP2']

  return(lr[['LocP1']])
}


lr_vector <- function(e1_1, e1_2){
  lr_delta = vector()
  for (counter in c(0:resolution)){
    useP1 = counter/resolution
    lr = LR(BayesNet = BayesNet, useP1 = useP1, useP2 =
        1-useP1, e1_1 = e1_1, e1_2 = e1_2)
    lr_delta <- rbind(lr_delta, c(useP1, lr))
  }
  return(lr_delta)
}


plot(lr_vector(1, 0.1), xlim = c(0,1),ylim=c(0,120),
   xlab = 'Pr(UseP1)', ylab = 'LR', type = 'l', main =
   'LR on person-level (varying Pr(UseP1))')


lines(lr_vector(1,0.01), lty = 2)


lines(lr_vector(1,0.001), lty = 4)

abline(h=1, lty = 3)
```

```
legend(x=0,y=120,cex=0.7,legend = c('LR_dev␣=␣10', 'LR_
    dev␣=␣100', 'LR_dev␣=␣1000'), lty=c(1,2,4))
```

B.2 Influence of $Pr(UseP_1)$

```
setwd("C:/Users/Hannes/switchdrive/Thesis")

library(RHugin)

BayesNet <- read.rhd("BN_Reduced_Sim2.net")




LR <- function(BayesNet, useP1, useP2, e1_1, e1_2, e2_
    1, e2_2){

  tabUseP <- get.table(BayesNet, 'UseP')
  tabUseP[1,'Freq'] = useP1
  tabUseP[2,'Freq'] = useP2
  tabUseP[3,'Freq'] = 1-useP1-useP2
  set.table(BayesNet, 'UseP', tabUseP)

  tabE1 <- get.table(BayesNet, 'E1')
  tabE1[1,'Freq'] = e1_1
  tabE1[2,'Freq'] = 1-e1_1
  tabE1[3,'Freq'] = e1_2
  tabE1[4,'Freq'] = 1-e1_2
  set.table(BayesNet, 'E1', tabE1)

  tabE2 <- get.table(BayesNet, 'E2')
  tabE2[1,'Freq'] = e2_1
  tabE2[2,'Freq'] = 1-e2_1
  tabE2[3,'Freq'] = e2_2
  tabE2[4,'Freq'] = 1-e2_2
  set.table(BayesNet, 'E2', tabE2)

  compile(BayesNet)
  set.finding(BayesNet, 'E1', 'E1')
```

```
  set.finding(BayesNet, 'E2', 'E2')
  propagate(BayesNet)
  odds <- get.belief(BayesNet, 'LocP')
  uncompile(BayesNet)

  lr <- odds['LocP1'] / odds['LocP2']

  return(lr[['LocP1']])
}


lr_vector <- function(e1_1, e1_2, e2_1, e2_2){
  lr_delta = vector()
  for (counter in c(0:resolution)){
    useP1 = counter/resolution
    lr = LR(BayesNet = BayesNet, useP1 = useP1, useP2 =
        1-useP1, e1_1 = e1_1, e1_2 = e1_2, e2_1 = e2_1,
        e2_2 = e2_2)
    lr_delta <- rbind(lr_delta, c(useP1, lr))
  }
  return(lr_delta)
}

resolution = 100

sim_run <- function(e2_1, e2_2, y_lim){
  LR_P = e2_1 / e2_2
  title = paste('LR on person-level \n(varying Pr(UseP1
     ), LR(E2)=', LR_P, ')')
  plot(lr_vector(1, 0.1, e2_1, e2_2), xlim = c(0,1),
     ylim=c(0,y_lim), xlab = 'Pr(UseP1)', ylab = 'LR',
     type = 'l', main = title)
  lines(lr_vector(1, 0.01, e2_1, e2_2), lty = 2)
  lines(lr_vector(1, 0.001, e2_1, e2_2), lty = 4)
  abline(h=1, lty = 3)
  legend(x=0,y=y_lim,cex=0.7,legend = c('LR_dev = 10',
     'LR_dev = 100', 'LR_dev = 1000'), lty=c(1,2,4))
}


sim_run(1, 0.1, 150)
```

```
sim_run(1, 0.01, 1100)
sim_run(1, 0.001, 1100)
```

B.3 Influence of α and β

```
setwd("C:/Users/Hannes/switchdrive/Thesis")

library(RHugin)

BayesNet <- read.rhd("BN_Reduced_Sim3.net")




LR <- function(BayesNet, alpha, beta, gama, delta,
    theta, e1_1, e1_2, e3_1, e3_2, e4_1, e4_2, P_user1,
    P_user2){

  tabUseP <- get.table(BayesNet, 'UseP')
  tabUseP[1,'Freq'] = alpha
  tabUseP[2,'Freq'] = beta
  tabUseP[3,'Freq'] = 1-alpha-beta
  tabUseP[4,'Freq'] = gama
  tabUseP[5,'Freq'] = delta
  tabUseP[6,'Freq'] = 1-gama-delta
  set.table(BayesNet, 'UseP', tabUseP)

  tabE1 <- get.table(BayesNet, 'E1')
  tabE1[1,'Freq'] = e1_1
  tabE1[2,'Freq'] = 1-e1_1
  tabE1[3,'Freq'] = e1_2
  tabE1[4,'Freq'] = 1-e1_2
  set.table(BayesNet, 'E1', tabE1)

  tabE3 <- get.table(BayesNet, 'E3')
  tabE3[1,'Freq'] = e3_1
  tabE3[2,'Freq'] = 1-e3_1
  tabE3[3,'Freq'] = e3_2
  tabE3[4,'Freq'] = 1-e3_2
```

```
  set.table(BayesNet, 'E3', tabE3)

  tabE4 <- get.table(BayesNet, 'E4')
  tabE4[1,'Freq'] = e4_1
  tabE4[2,'Freq'] = 1-e4_1
  tabE4[3,'Freq'] = e4_2
  tabE4[4,'Freq'] = 1-e4_2
  set.table(BayesNet, 'E4', tabE4)

  tabUser <- get.table(BayesNet, 'User')
  tabUser[1,'Freq'] = P_user1
  tabUser[2,'Freq'] = P_user2
  tabUser[3,'Freq'] = 1 - P_user1 -P_user2

  tabUseU <- get.table(BayesNet, 'UseU')
  tabUseU[13,'Freq'] = theta
  tabUseU[14,'Freq'] = 1-theta

  compile(BayesNet)
  set.finding(BayesNet, 'E1', 'E1')
  set.finding(BayesNet, 'E3', 'E3')
  set.finding(BayesNet, 'E4', 'E4')
  propagate(BayesNet)
  odds <- get.belief(BayesNet, 'LocP')
  uncompile(BayesNet)

  lr <- odds['LocP1'] / odds['LocP2']

  return(lr[['LocP1']])
}


resolution = 100

e1_1 = 1
e1_2 = 0.001
e3_1 = 1
e3_2 = 0.001
e4_1 = 1
e4_2 = 0.001
useP1 = 0.5
```

```
theta = 0.8

lr_vector <- function(gama){
  lr_vec = vector()
  for (counter in c(0:resolution)){
    y_val = counter/resolution
    lr = LR(BayesNet = BayesNet, alpha = y_val, beta =
        (1 - y_val) * fac_bet, gama = gama, delta = (1 -
         gama) * fac_delta, P_user1 = useP1, P_user2 =
        1-useP1, e1_1 = e1_1, e1_2 = e1_2, e3_1 = e3_1,
        e3_2 = e3_2, e4_1 = e4_1, e4_2=e4_2, theta =
        theta)
    lr_vec <- rbind(lr_vec, c(y_val, lr))
  }
  return(lr_vec)
}

ylim_top = 1300

fac_bet = 1
fac_delta = 1
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'alpha', ylab = 'LR', type = 'l', main = '
   alpha varied, gamma fixed, beta = 1 - alpha')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('gamma = 0.9',
     'gamma = 0.5', 'gamma = 0.1'), lty=c(1,2,4))


fac_bet=0.9
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'alpha', ylab = 'LR', type = 'l', main = '
   alpha varied, gamma fixed, beta = 0.9 (1 - alpha)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('gamma = 0.9',
     'gamma = 0.5', 'gamma = 0.1'), lty=c(1,2,4))
```

```
fac_bet=0.5
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'alpha', ylab = 'LR', type = 'l', main = '
   alpha varied, gamma fixed, beta = 0.5 (1 - alpha)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('gamma = 0.9',
    'gamma = 0.5', 'gamma = 0.1'), lty=c(1,2,4))

fac_bet=0.1
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'alpha', ylab = 'LR', type = 'l', main = '
   alpha varied, gamma fixed, beta = 0.1(1 - alpha)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('gamma = 0.9',
    'gamma = 0.5', 'gamma = 0.1'), lty=c(1,2,4))
```

B.4 Influence of γ and δ

```
setwd("C:/Users/Hannes/switchdrive/Thesis")

library(RHugin)

BayesNet <- read.rhd("BN_Reduced_Sim3.net")

LR <- function(BayesNet, alpha, beta, gama, delta,
   theta, e1_1, e1_2, e3_1, e3_2, e4_1, e4_2, P_user1,
   P_user2){

  tabUseP <- get.table(BayesNet, 'UseP')
  tabUseP[1,'Freq'] = alpha
  tabUseP[2,'Freq'] = beta
  tabUseP[3,'Freq'] = 1-alpha-beta
  tabUseP[4,'Freq'] = gama
  tabUseP[5,'Freq'] = delta
  tabUseP[6,'Freq'] = 1-gama-delta
```

```
set.table(BayesNet, 'UseP', tabUseP)

tabE1 <- get.table(BayesNet, 'E1')
tabE1[1,'Freq'] = e1_1
tabE1[2,'Freq'] = 1-e1_1
tabE1[3,'Freq'] = e1_2
tabE1[4,'Freq'] = 1-e1_2
set.table(BayesNet, 'E1', tabE1)

tabE3 <- get.table(BayesNet, 'E3')
tabE3[1,'Freq'] = e3_1
tabE3[2,'Freq'] = 1-e3_1
tabE3[3,'Freq'] = e3_2
tabE3[4,'Freq'] = 1-e3_2
set.table(BayesNet, 'E3', tabE3)

tabE4 <- get.table(BayesNet, 'E4')
tabE4[1,'Freq'] = e4_1
tabE4[2,'Freq'] = 1-e4_1
tabE4[3,'Freq'] = e4_2
tabE4[4,'Freq'] = 1-e4_2
set.table(BayesNet, 'E4', tabE4)

tabUser <- get.table(BayesNet, 'User')
tabUser[1,'Freq'] = P_user1
tabUser[2,'Freq'] = P_user2
tabUser[3,'Freq'] = 1 - P_user1 -P_user2

tabUseU <- get.table(BayesNet, 'UseU')
tabUseU[13,'Freq'] = theta
tabUseU[14,'Freq'] = 1-theta

compile(BayesNet)
set.finding(BayesNet, 'E1', 'E1')
set.finding(BayesNet, 'E3', 'E3')
set.finding(BayesNet, 'E4', 'E4')
propagate(BayesNet)
odds <- get.belief(BayesNet, 'LocP')
uncompile(BayesNet)

lr <- odds['LocP1'] / odds['LocP2']
```

```
  return(lr[['LocP1']])
}


resolution = 100

e1_1 = 1
e1_2 = 0.001
e3_1 = 1
e3_2 = 0.001
e4_1 = 1
e4_2 = 0.001
useP1 = 0.5
theta = 0.8

lr_vector <- function(alpha){
  lr_vec = vector()
  for (counter in c(0:resolution)){
    y_val = counter/resolution
    lr = LR(BayesNet = BayesNet, alpha = alpha, beta =
        (1 - alpha) * fac_bet, gama = y_val, delta = (1
        - y_val) * fac_delta, P_user1 = useP1, P_user2 =
         1-useP1, e1_1 = e1_1, e1_2 = e1_2, e3_1 = e3_1,
         e3_2 = e3_2, e4_1 = e4_1, e4_2=e4_2, theta =
        theta)
    lr_vec <- rbind(lr_vec, c(y_val, lr))
  }
  return(lr_vec)
}

ylim_top = 1400

fac_bet = 0.9
fac_delta = 1
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'gamma', ylab = 'LR', type = 'l', main = '
   gamma varied, alpha fixed, delta = 1 - gamma')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
```

```
legend(x=0,y=ylim_top,cex=0.6,legend = c('alpha = 0.9',
    'alpha = 0.5', 'alpha = 0.1'), lty=c(1,2,4))


fac_delta=0.9
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'gamma', ylab = 'LR', type = 'l', main = '
   gamma varied, alpha fixed, delta = 0.9(1 - gamma)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('alpha = 0.9',
    'alpha = 0.5', 'alpha = 0.1'), lty=c(1,2,4))


fac_delta=0.5
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'gamma', ylab = 'LR', type = 'l', main = '
   gamma varied, alpha fixed, delta = 0.5(1 - gamma)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('alpha = 0.9',
    'alpha = 0.5', 'alpha = 0.1'), lty=c(1,2,4))

fac_delta=0.1
plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'gamma', ylab = 'LR', type = 'l', main = '
   gamma varied, alpha fixed, delta = 0.1(1 - gamma)')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('alpha = 0.9',
    'alpha = 0.5', 'alpha = 0.1'), lty=c(1,2,4))
```

B.5 Influence of $Pr(User_1)$ and θ

```
setwd("C:/Users/Hannes/switchdrive/Thesis")
```

```
library(RHugin)

BayesNet <- read.rhd("BN_Reduced_Sim3.net")




LR <- function(BayesNet, alpha, beta, gama, delta,
   theta, e1_1, e1_2, e3_1, e3_2, e4_1, e4_2, P_user1,
   P_user2){

  tabUseP <- get.table(BayesNet, 'UseP')
  tabUseP[1,'Freq'] = alpha
  tabUseP[2,'Freq'] = beta
  tabUseP[3,'Freq'] = 1-alpha-beta
  tabUseP[4,'Freq'] = gama
  tabUseP[5,'Freq'] = delta
  tabUseP[6,'Freq'] = 1-gama-delta
  set.table(BayesNet, 'UseP', tabUseP)

  tabE1 <- get.table(BayesNet, 'E1')
  tabE1[1,'Freq'] = e1_1
  tabE1[2,'Freq'] = 1-e1_1
  tabE1[3,'Freq'] = e1_2
  tabE1[4,'Freq'] = 1-e1_2
  set.table(BayesNet, 'E1', tabE1)

  tabE3 <- get.table(BayesNet, 'E3')
  tabE3[1,'Freq'] = e3_1
  tabE3[2,'Freq'] = 1-e3_1
  tabE3[3,'Freq'] = e3_2
  tabE3[4,'Freq'] = 1-e3_2
  set.table(BayesNet, 'E3', tabE3)

  tabE4 <- get.table(BayesNet, 'E4')
  tabE4[1,'Freq'] = e4_1
  tabE4[2,'Freq'] = 1-e4_1
  tabE4[3,'Freq'] = e4_2
  tabE4[4,'Freq'] = 1-e4_2
  set.table(BayesNet, 'E4', tabE4)
```

```
  tabUser <- get.table(BayesNet, 'User')
  tabUser[1,'Freq'] = P_user1
  tabUser[2,'Freq'] = P_user2
  tabUser[3,'Freq'] = 1 - P_user1 -P_user2

  tabUseU <- get.table(BayesNet, 'UseU')
  tabUseU[13,'Freq'] = theta
  tabUseU[14,'Freq'] = 1-theta

  compile(BayesNet)
  set.finding(BayesNet, 'E1', 'E1')
  set.finding(BayesNet, 'E3', 'E3')
  set.finding(BayesNet, 'E4', 'E4')
  propagate(BayesNet)
  odds <- get.belief(BayesNet, 'LocP')
  uncompile(BayesNet)

  lr <- odds['LocP1'] / odds['LocP2']

  return(lr[['LocP1']])
}


resolution = 100

e1_1 = 1
e1_2 = 0.001
e3_1 = 1
e3_2 = 0.001
e4_1 = 1
e4_2 = 0.001
alpha = 0.5
beta = 0.3
gama = 0.2
delta = 0.5


lr_vector <- function(theta){
  lr_vec = vector()
  for (counter in c(0:resolution)){
    y_val = counter/resolution
```

```
    lr = LR(BayesNet = BayesNet, alpha = alpha, beta =
      beta, gama = gama, delta = delta, P_user1 =
      resolution, P_user2 = 1-resolution, e1_1 = e1_1,
       e1_2 = e1_2, e3_1 = e3_1, e3_2 = e3_2, e4_1 =
      e4_1, e4_2=e4_2, theta = theta)
    lr_vec <- rbind(lr_vec, c(y_val, lr))
  }
  return(lr_vec)
}

ylim_top = 1300


e3_1 = 1
e3_2 = 0.001
e4_1 = 1
e4_2 = 0.001

plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'P(User1)', ylab = 'LR', type = 'l', main = '
   P(User1) varied, theta fixed')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('theta = 0.9',
    'theta = 0.5', 'theta = 0.1'), lty=c(1,2,4))

e3_1 = 1
e3_2 = 0.01
e4_1 = 1
e4_2 = 0.01

plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'P(User1)', ylab = 'LR', type = 'l', main = '
   P(User1) varied, theta fixed')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('theta = 0.9',
    'theta = 0.5', 'theta = 0.1'), lty=c(1,2,4))
```

```
e3_1 = 1
e3_2 = 0.5
e4_1 = 1
e4_2 = 0.5

plot(lr_vector(0.9), xlim = c(0,1),ylim=c(0,ylim_top),
   xlab = 'P(User1)', ylab = 'LR', type = 'l', main = '
   P(User1) varied, theta fixed')
lines(lr_vector(0.5), lty = 2)
lines(lr_vector(0.1), lty = 4)
abline(h=1, lty = 3)
legend(x=0,y=ylim_top,cex=0.6,legend = c('theta = 0.9',
     'theta = 0.5', 'theta = 0.1'), lty=c(1,2,4))
```

Appendix C

Sc2 & 4: List of behavioural biometric characteristics

In the following a full list of the observed event-categories used in chapters 5 and 7 as indicators for user-behaviour is presented. They are reused as such from (Michelet, 2021).

Each event is in one of three categories:

Punctual: Punctual events happen from time to time. For each punctual event-type, the number of events is counted and the total duration of these events as well as the average event duration are calculated.

Non-Punctual: Non-punctual events are constant and have one of two states: off (0) and on (1). At any given moment in time, the phone is in either one of those two states. For each state, the number of events in this state is counted and the total duration and the average duration per event are calculated, resulting in 6 variables per event-type.

Specific: Specific event-types are treated in their own individualised manner:

- Notifications: The number of events is calculated for each of the following events:
  - «Hidden»
  - «Dismiss»
  - «IndirectClear»
  - «Recieve»

  - «Orb»
  - «DefaultAction»

- Battery Percentage: Total amount of battery percents used.
- Siri: Number of Siri usages
- App usage & in Focus: The day is split up in 12 sessions of 2h length. For each session, events in this event-type are treated as punctual events.
- Lockdown: Number of power-on events of the device

name	type	description
/display/orientation	non-punctual	Orientation of the screen (portrait / landscape)
/device/isPluggedIn	non-punctual	Is the device currently plugged in
/display/isBacklit	non-punctual	Back-lighting of the screen
/device/isLocked	non-punctual	Current lock-state of the device
/system/airplaneMode	non-punctual	State of the airplane mode
/wifi/connexion	punctual	established WiFi connections
/bluetooth/isConnected	non-punctual	Current bluetooth state
/device/batterySaver	punctual	Battery save mode turned on
/audio/outputRoute	non-punctual	State of audio output
/audio/inputRoute	non-punctual	State of audio input
/notification/usage	specific	Notification events
/device/lowPowerMode	non-punctual	State of low power mode
/device/batteryPercentage	specific	Current battery percentage
/siri/ui	specific	Siri Events
/media/nowPlaying	non-punctual	State of media play
/app/usage	specific / punctual	Application usage
/app/inFocus	specific / punctual	Application usage in foreground
lockdown.log	specific	Power-off events of the device

Appendix D

Sc2 & 4: Code used for Feature Extraction

This Python script was used in Chapters 5 and 7 to import and anonymize characteristics from the knowledgeC.db and the lockdown.log. This script is mostly adapted from (Michelet, 2021).

The script is available at `https://github.com/HSpichig/Thesis/blob/main/BB_anonimize.py`

```
#imports des diff rents modules utiles au projet
import pandas as pd
import sqlite3
from datetime import *
import re
import os
import subprocess
import numpy as np

#Cette fonction permet de cr er un dataframe contenant
    toutes les dates allant de la date du d but (Start
  )   la date de fin (End)
def create_default_dataframe(Start,End) :
    #cr ation du datetime pour la date de d part
    x = datetime.strptime(Start.strftime("%d/%m/%y ") +
        "13:00:00","%d/%m/%y %H:%M:%S")
    date_list = []
    #boucle ajoutant la date pour chaque jour contenu
       entre la date de d part et la date d'arriv e
    while x.date() <= End :
```

```
        date_list.append(x.date())
        x += timedelta(hours=24)

    #cr ation d'un dataframe    partir de cette liste
       de dates
    df_dict = {"Date" : date_list}
    df = pd.DataFrame(df_dict)

    return df


#fonction permettant de cr er plusieurs    vnements
   se d roulant chacun sur une journ e lorsque l'un
   des    vnements    se d roule sur plusieurs jours
def set_time_and_sort(df):
    #compteur
    c = 1
    #parcourt le dataframe et supprime les anomalies (
       timestamp plus petits que 0)
    for x in df.index:
        if df.loc[x]["Start_ts"] < 0 or df.loc[x]["
           End_ts"] < 0:
            df = df.drop(index=x, axis=0)
            print(f"ignoring line {x}")
    #   chaque fois
    while True :
        #indique le num ro de passage
        print("passage n " + str(c))
        c += 1
        test = False
        l1 = []
        c2 = 0
        counter = 0
        #parcourt tous les    lments    du dataframe
        for x in df.index:
            #si un    vnement    a lieu sur plusieurs
               jours et que le compteur est plus petit
               que 31
            if datetime.fromtimestamp(df.loc[x]["
               Start_ts"]).date() != datetime.
               fromtimestamp(
```

```
    df.loc[x]["End_ts"]).date() and c
      <= 31:
test = True
#copie les  lments  relatifs  cet
     vnement  deux fois
temp_dic_1 = {index: df.loc[x][index]
   for index in df.loc[x].index}
temp_dic_2 = {index: df.loc[x][index]
   for index in df.loc[x].index}
#modifie la date de fin de la premi re
    copie pour qu'elle se termine le
   m me jour que la date de
   commencement, mais  23:59:59
End_dt_1 = datetime.strptime(str(
   datetime.fromtimestamp(temp_dic_1["
   Start_ts"]).date()) + " 23:59:59",
                                "%Y-%m-%d
                                  %H:%M:%
                                  S")
# modifie la date de commencement de la
    deuxi me copie. Cette date sera le
    jour suivant le jour de
   commencement, mais  00:00:00
Start_temp = datetime.fromtimestamp(
   temp_dic_2["Start_ts"]) + timedelta(
   days=1)
Start_dt_2 = datetime.strptime(str(
   Start_temp.date()) + " 00:00:00",
                                  "%Y-%m-%
                                    d %H
                                    :%M:%
                                    S")
#r calcule la dur e des  vnements
temp_dic_1["End_ts"] = int(datetime.
   timestamp(End_dt_1))
temp_dic_1["Duration"] = temp_dic_1["
   End_ts"] - (temp_dic_1["Start_ts"])
temp_dic_2["Start_ts"] = int(datetime.
   timestamp(Start_dt_2))
temp_dic_2["Duration"] = (temp_dic_2["
   End_ts"]) - temp_dic_2["Start_ts"]
```

```
                #ajoute les deux   lments   fraichement
                    cr   s   une liste , et supprimme
                   l 'ancien
                l1 . append (temp_dic_1)
                l1 . append (temp_dic_2)
                df = df . drop (index=x , axis=0)
                c2 += 1
        print ( f "{c2}␣elements␣were␣cleaned" )

        #s ' il n 'y a plus d '   lments     traiter ou que
            le compteur d passe 31, sort de la boucle
        if test == False :
            print ( "No␣more␣element␣to␣clean" )
            break
        #ajoutes les   vnements   nouvellement cr   s
           au dataframe avant de recommencer la boucle
        d = {key : [] for key in l1 [0]. keys ()}
        for dico in l1 :
            for key in dico . keys () :
                d [ key ] . append ( dico [ key ] )

        df = pd . concat ( [ pd . DataFrame (d) , df ] ) .
           sort_values (by=[ "Start_ts" ] ) . reset_index ()
        df = df . drop (columns="index" )

    #lorsque la boucle est finie , calcule les champs
       manquants n cessaires
    df = df . assign (
        #date et heure de commencement
        Start_Date=list (map(lambda x : datetime .
           fromtimestamp ( int (x) ) . date () , df . Start_ts ) ) ,
        Start_Time=list (map(lambda x : datetime .
           fromtimestamp ( int (x) ) . time () , df . Start_ts ) ) ,
        #date et heure de fin
        End_Date=list (map(lambda x : datetime .
           fromtimestamp ( int (x) ) . date () , df . End_ts ) ) ,
        End_Time=list (map(lambda x : datetime .
           fromtimestamp ( int (x) ) . time () , df . End_ts ) ) ,
        #datetime de commencement et de fin
        Start_dt=list (map(lambda x : datetime .
           fromtimestamp ( int (x) ) , df . Start_ts ) ) ,
```

```
        End_dt=list(map(lambda x: datetime.
            fromtimestamp(int(x)), df.End_ts)),
        #champ utile    l'agr gation
        Count=lambda y: [1 for _ in df.Z_PK]
    )
    #retourne le dataframe nettoy
    return df

#fonction permettant de contr ler si un   vnement   a
   eu lieu durant une session ou non (il n'appartient
   pas    une session si le d but et la fin de l'
     vnement   se situent tous deux avant le d but de
   la session ou apr s la fin de la session
def is_in_session(start,end,time):
    return 0 if ((start<datetime.strptime(start.
        strftime("%Y-%m-%d ") + time[0],"%Y-%m-%d %H:%M
        :%S") and end<datetime.strptime(end.strftime("%Y
        -%m-%d ") + time[0],"%Y-%m-%d %H:%M:%S")) or (
        start>datetime.strptime(start.strftime("%Y-%m-%d
         ") + time[1],"%Y-%m-%d %H:%M:%S") and end>
        datetime.strptime(end.strftime("%Y-%m-%d ") +
        time[1],"%Y-%m-%d %H:%M:%S"))) else 1

#fonction permettant de cr er les diff rentes
   sessions d'utilisation du t l phone
def set_sessions(df) :
    l = []
    #cr e 12 sessions de 2h (00:00:00-01:59:59 /
        02:00:00-03:59:59 / ...)
    for x in range(12) :
        if x >= 5 :
            l.append((str(2*x) + ":00:00",str(2*x+1) +
                ":59:59"))
        else :
            l.append(("0" + str(2*x) + ":00:00","0" +
                str(2*x+1) + ":59:59","%H:%M:%S"))

    #pour chacune des sessions, contr le si l'
          vnement   a eu lieu durant la session (1) ou
       non (0)
    df = df.assign(
```

```
        Session_0=list(map(lambda x,y: is_in_session(x,
            y,l[0]) ,df.Start_dt,df.End_dt)),
        Session_1=list(map(lambda x,y: is_in_session(x,
            y,l[1]) ,df.Start_dt,df.End_dt)),
        Session_2=list(map(lambda x,y: is_in_session(x,
            y,l[2]) ,df.Start_dt,df.End_dt)),
        Session_3=list(map(lambda x,y: is_in_session(x,
            y,l[3]) ,df.Start_dt,df.End_dt)),
        Session_4=list(map(lambda x,y: is_in_session(x,
            y,l[4]) ,df.Start_dt,df.End_dt)),
        Session_5=list(map(lambda x,y: is_in_session(x,
            y,l[5]) ,df.Start_dt,df.End_dt)),
        Session_6=list(map(lambda x,y: is_in_session(x,
            y,l[6]) ,df.Start_dt,df.End_dt)),
        Session_7=list(map(lambda x,y: is_in_session(x,
            y,l[7]) ,df.Start_dt,df.End_dt)),
        Session_8=list(map(lambda x,y: is_in_session(x,
            y,l[8]) ,df.Start_dt,df.End_dt)),
        Session_9=list(map(lambda x,y: is_in_session(x,
            y,l[9]) ,df.Start_dt,df.End_dt)),
        Session_10=list(map(lambda x,y: is_in_session(x
            ,y,l[10]) ,df.Start_dt,df.End_dt)),
        Session_11=list(map(lambda x,y: is_in_session(x
            ,y,l[11]) ,df.Start_dt,df.End_dt))
    )

    #retourne le dataframe
    return df

#fonction permettant de cr er le datframe de base pour
    les   vnements   non ponctuels
def get_default_df_knowledge_non_ponctual(Start,End,
   name):
    #r cup re le dataframe avec tous les jours d'
        utilisation
    df = create_default_dataframe(Start, End)

    #cr     les colonnes de 0
    df = df.assign(col_1=[0 for x in df.Date],
                   col_2=[0 for x in df.Date],
                   col_3=[0 for x in df.Date],
```

```
                col_4=[0 for x in df.Date],
                col_5=[0 for x in df.Date],
                col_6=[0 for x in df.Date])

    #renomme les colonnes sous le format :
        nom_ v nement + nom_variable
    df = df.rename(
        columns={"col_1": name + "_1_nb", "col_2": name
            + "_0_nb",
                "col_3": name + "_1_duration",
                "col_4": name + "_0_duration", "col_5"
                    : name + "_1_duration_mean",
                "col_6": name + "_0_duration_mean"})

    #Met la date en index et retourne le dataframe
    df = df.rename(columns={"Date": "Start_Date"})
    df = df.set_index("Start_Date")

    return df

#fonction permettant d'agr ger les donn es des
      vnements   non ponctuels
def anonimize_knowledge_dataframe_non_ponctual_events(
    df,name):

    #nombre de position on
    is_on = df.loc[(df.ValueDouble != 0.0) & (df.
        ValueDouble != 2.0)][["Start_Date", "Count"]].
        groupby("Start_Date").sum("Count")
    is_on = is_on.rename(columns={"Count": name + "
        _1_nb"})

    # nombre de position off
    is_off = df.loc[(df.ValueDouble == 0.0) | (df.
        ValueDouble == 2.0)][["Start_Date", "Count"]].
        groupby("Start_Date").sum("Count")
    is_off = is_off.rename(columns={"Count": name + "
        _0_nb"})

    # dur e totale en position on
    is_on_duration = df.loc[(df.ValueDouble == 1.0)][["
```

```
    Start_Date", "Duration"]].groupby("Start_Date").
    sum(
     "Duration")
is_on_duration = is_on_duration.rename(columns={"
    Duration": name + "_1_duration"})

# dur e totale en position off
is_off_duration = df.loc[(df.ValueDouble == 0.0)][[
    "Start_Date", "Duration"]].groupby("Start_Date")
    .sum(
     "Duration")
is_off_duration = is_off_duration.rename(columns={"
    Duration": name + "_0_duration"})

# dur e moyenne en position on
is_on_duration_mean = df.loc[(df.ValueDouble ==
    1.0)][["Start_Date", "Duration"]].groupby("
    Start_Date").mean(
     "Duration")
is_on_duration_mean = is_on_duration_mean.rename(
    columns={"Duration": name + "_1_duration_mean"})

# dur e moyenne en position off
is_off_duration_mean = df.loc[(df.ValueDouble ==
    0.0)][["Start_Date", "Duration"]].groupby("
    Start_Date").mean(
     "Duration")
is_off_duration_mean = is_off_duration_mean.rename(
    columns={"Duration": name + "_0_duration_mean"})

#regroupement des dataframe, li s par la date qui
    est devenue l'index via les groupby
merge = pd.concat(
     [is_on,is_off,is_on_duration,is_off_duration,
        is_on_duration_mean,is_off_duration_mean],
     axis=1)
#remplissage des   lments   nuls avec des 0 et
    retour du dataframe
merge = merge.fillna(0)
return merge
```

```
#fonction permettant de cr er le dataframe de base
    pour les    lments    ponctuels
def get_default_df_knowledge_ponctual(Start,End,name):
    #cr ation du dataframe avec les jours d'
        utilisation
    df = create_default_dataframe(Start, End)

    #cr ation des colonnes de 0
    df = df.assign(col_1=[0 for x in df.Date],
                   col_2=[0 for x in df.Date],
                   col_3=[0 for x in df.Date])

    #renomme les colonnes au format : nom_ v nement +
        nom_variable
    df = df.rename(
        columns={"col_1": name + "_event_nb", "col_2":
           name + "_event_duration",
                 "col_3": name + "_event_duration_mean"
                    })

    #Met la date en index et retourne le dataframe
    df = df.rename(columns={"Date":"Start_Date"})
    df = df.set_index("Start_Date")

    return df

#fonction permettant d'agr ger les donn es relatives
    aux    vnements    ponctuels
def anonimize_knowledge_dataframe_ponctual_events(df,
    name):
    #nombre d'    vnements
    event = df.loc[(df.ValueDouble != 0.0)][["
        Start_Date", "Count"]].groupby("Start_Date").sum
        ("Count")
    event = event.rename(columns={"Count": name + "
        _event_nb"})

    #dur e totale des    vnements
    event_duration = df.loc[(df.ValueDouble != 0.0)][["
        Start_Date", "Duration"]].groupby("Start_Date").
        sum(
```

```
        "Duration")
    event_duration = event_duration.rename(columns={"
       Duration": name + "_event_duration"})

    #dur e moyenne des   vnements
    event_duration_mean = df.loc[(df.ValueDouble !=
       0.0)][["Start_Date", "Duration"]].groupby("
       Start_Date").mean(
        "Duration")
    event_duration_mean = event_duration_mean.rename(
       columns={"Duration": name + "
       _event_duration_mean"})

    #fusion des dataframe via la date (devenue index
       avec les groupby)
    merge = pd.concat(
        [event,event_duration,event_duration_mean],
        axis=1)

    #remplissage des   lments   nuls avec des 0 et
       retour du dataframe
    merge = merge.fillna(0)
    return merge

#cr ation du dataframe de base pour les   lments
   notification
def get_default_df_knowledge_notification(Start,End):
    #r cup ration du dataframe avec les jours d'
       utilisation
    df = create_default_dataframe(Start, End)

    #cr ation des colonnes de 0 (nomm es correctement
        car les 6 noms sont connus)
    df = df.assign(Hidden=[0 for x in df.Date],
                   Receive=[0 for x in df.Date],
                   Dismiss=[0 for x in df.Date],
                   Orb=[0 for x in df.Date],
                   IndirectClear=[0 for x in df.Date],
                   DefaultAction=[0 for x in df.Date])

    #La date est mise en index et le dataframe est
```

```
        retourn
    df = df.rename(columns={"Date": "Start_Date"})
    df = df.set_index("Start_Date")

    return df

#fonction permettant d'agr ger les donn es li es aux
        vnements notifications
def anonimize_knowledge_notification(df) :

    #nombre de notifications hidden
    Hidden = df.loc[(df.Value == "Hidden")][["
        Start_Date", "Count"]].groupby("Start_Date").sum
        ("Count")
    Hidden = Hidden.rename(columns={"Count": "Hidden_nb
        "})

    # nombre de notifications receive
    Receive = df.loc[(df.Value == "Receive")][["
        Start_Date", "Count"]].groupby("Start_Date").sum
        ("Count")
    Receive = Receive.rename(columns={"Count": "
        Receive_nb"})

    # nombre de notifications dismiss
    Dismiss = df.loc[(df.Value == "Dismiss")][["
        Start_Date", "Count"]].groupby("Start_Date").sum
        ("Count")
    Dismiss = Dismiss.rename(columns={"Count": "
        Dismiss_nb"})

    # nombre de notifications orb
    Orb = df.loc[(df.Value == "Orb")][["Start_Date", "
        Count"]].groupby("Start_Date").sum("Count")
    Orb = Orb.rename(columns={"Count": "Orb_nb"})

    # nombre de notifications indirectclear
    IndirectClear = df.loc[(df.Value == "IndirectClear"
        )][["Start_Date", "Count"]].groupby("Start_Date"
        ).sum("Count")
    IndirectClear = IndirectClear.rename(columns={"
```

```
        Count": "IndirectClear_nb"})

    # nombre de notifications defaultaction
    DefaultAction = df.loc[(df.Value == "DefaultAction"
        )][["Start_Date", "Count"]].groupby("Start_Date"
        ).sum("Count")
    DefaultAction = DefaultAction.rename(columns={"
        Count": "DefaultAction_nb"})

    #fusion des dataframe via la date (mise en index
        par le groupby)
    merge = pd.concat(
        [Hidden,Receive,Dismiss,Orb,IndirectClear,
            DefaultAction],
        axis=1)

    #remplissage des    lments    nuls avec des 0 et
        retour du dataframe
    merge = merge.fillna(0)
    return merge

#fonction permettant de cr er le dataframe de base
    pour les donn es batterie et siri
def get_default_df_knowledge_percentage_siri(Start,End,
    name):
    #r cup ration du dataframe compos  des jours d'
        utilisation
    df = create_default_dataframe(Start, End)

    #cr ation de la colonne de 0
    df = df.assign(col_1=[0 for x in df.Date])

    #renommage de la colone (nom_ v nement +
        nom_variable)
    df = df.rename(
        columns={"col_1": name + "_nb"})

    #Mise en index de la date et retour du dataframe
    df = df.rename(columns={"Date": "Start_Date"})
    df = df.set_index("Start_Date")
```

```
    return df

#fonction permettant d'anonimiser les donn es li es
   aux  vnements  batterie et siri
def anonimize_percentage_and_siri(df, name) :

    #nombre d'occurences de l'  vnement
    Count = df[["Start_Date", "Count"]].groupby("
        Start_Date").sum("Count")
    Count = Count.rename(columns={"Count": name + "_nb"
        })

    #remplissage des valeurs nulles avec des 0 et
        retour du dataframe
    Count = Count.fillna(0)

    return Count

#fonction permettant de r cup rer le datframe de base
    pour les  vnements  li s  l'utilisation d'
   applications
def get_default_df_knowledge_app_usage(Start,End,name)
   :
    #r cup ration du dataframe contenant les jours d'
        utilisation
    df = create_default_dataframe(Start, End)

    #cr ation des colonnes de 0
    df = df.assign(col_1=[0 for x in df.Date],
                   col_2=[0 for x in df.Date],
                   col_3=[0 for x in df.Date],
                   col_4=[0 for x in df.Date],
                   col_5=[0 for x in df.Date],
                   col_6=[0 for x in df.Date],
                   col_7=[0 for x in df.Date],
                   col_8=[0 for x in df.Date],
                   col_9=[0 for x in df.Date],
                   col_10=[0 for x in df.Date],
                   col_11=[0 for x in df.Date],
                   col_12=[0 for x in df.Date])
```

```
    #renommage des colonnes (nom_ v nement +
       nom_variable)
    df = df.rename(
        columns={"col_1": name + "_Session_0", "col_2":
            name + "_Session_1", "col_3": name + "
           _Session_2",
                 "col_4": name + "_Session_3", "col_5":
                     name + "_Session_4",
                 "col_6": name + "_Session_5",
                 "col_7": name + "_Session_6", "col_8":
                     name + "_Session_7",
                 "col_9": name + "_Session_8",
                 "col_10": name + "_Session_9", "col_11
                    ": name + "_Session_10","col_12":
                    name + "_Session_11"})

    #Mise en index de la date et retour du dataframe
    df = df.rename(columns={"Date": "Start_Date"})
    df = df.set_index("Start_Date")

    return df

#fonction permettant d'agr ger les donn es li es aux
      vnements   d'utilisation des applications
def anonimize_app_usage(df,name) :
    #utilisation du t l phone dans la session 0
    Session_0 = df[["Start_Date","Session_0"]].groupby(
       "Start_Date").sum("Count")
    Session_0 = Session_0.assign(Session_0=list(map(
       lambda x : 1 if int(x)>0 else 0,Session_0.
       Session_0)))
    Session_0 = Session_0.rename(columns={"Session_0" :
        name + "_Session_0"})

    # utilisation du t l phone dans la session 1
    Session_1 = df[["Start_Date", "Session_1"]].groupby
       ("Start_Date").sum("Count")
    Session_1 = Session_1.assign(Session_1=list(map(
       lambda x: 1 if int(x) > 0 else 0, Session_1.
       Session_1)))
    Session_1 = Session_1.rename(columns={"Session_1":
```

```
    name + "_Session_1"})

# utilisation du t l phone dans la session 2
Session_2 = df[["Start_Date", "Session_2"]].groupby
    ("Start_Date").sum("Count")
Session_2 = Session_2.assign(Session_2=list(map(
    lambda x: 1 if int(x) > 0 else 0, Session_2.
    Session_2)))
Session_2 = Session_2.rename(columns={"Session_2":
    name + "_Session_2"})

# utilisation du t l phone dans la session 3
Session_3 = df[["Start_Date", "Session_3"]].groupby
    ("Start_Date").sum("Count")
Session_3 = Session_3.assign(Session_3=list(map(
    lambda x: 1 if int(x) > 0 else 0, Session_3.
    Session_3)))
Session_3 = Session_3.rename(columns={"Session_3":
    name + "_Session_3"})

# utilisation du t l phone dans la session 4
Session_4 = df[["Start_Date", "Session_4"]].groupby
    ("Start_Date").sum("Count")
Session_4 = Session_4.assign(Session_4=list(map(
    lambda x: 1 if int(x) > 0 else 0, Session_4.
    Session_4)))
Session_4 = Session_4.rename(columns={"Session_4":
    name + "_Session_4"})

# utilisation du t l phone dans la session 5
Session_5 = df[["Start_Date", "Session_5"]].groupby
    ("Start_Date").sum("Count")
Session_5 = Session_5.assign(Session_5=list(map(
    lambda x: 1 if int(x) > 0 else 0, Session_5.
    Session_5)))
Session_5 = Session_5.rename(columns={"Session_5":
    name + "_Session_5"})

# utilisation du t l phone dans la session 6
Session_6 = df[["Start_Date", "Session_6"]].groupby
    ("Start_Date").sum("Count")
```

```
Session_6 = Session_6.assign(Session_6=list(map(lambda x: 1 if int(x) > 0 else 0, Session_6.Session_6)))
Session_6 = Session_6.rename(columns={"Session_6": name + "_Session_6"})

# utilisation du t l phone dans la session 7
Session_7 = df[["Start_Date", "Session_7"]].groupby("Start_Date").sum("Count")
Session_7 = Session_7.assign(Session_7=list(map(lambda x: 1 if int(x) > 0 else 0, Session_7.Session_7)))
Session_7 = Session_7.rename(columns={"Session_7": name + "_Session_7"})

# utilisation du t l phone dans la session 8
Session_8 = df[["Start_Date", "Session_8"]].groupby("Start_Date").sum("Count")
Session_8 = Session_8.assign(Session_8=list(map(lambda x: 1 if int(x) > 0 else 0, Session_8.Session_8)))
Session_8 = Session_8.rename(columns={"Session_8": name + "_Session_8"})

# utilisation du t l phone dans la session 9
Session_9 = df[["Start_Date", "Session_9"]].groupby("Start_Date").sum("Count")
Session_9 = Session_9.assign(Session_9=list(map(lambda x: 1 if int(x) > 0 else 0, Session_9.Session_9)))
Session_9 = Session_9.rename(columns={"Session_9": name + "_Session_9"})

# utilisation du t l phone dans la session 10
Session_10 = df[["Start_Date", "Session_10"]].groupby("Start_Date").sum("Count")
Session_10 = Session_10.assign(Session_10=list(map(lambda x: 1 if int(x) > 0 else 0, Session_10.Session_10)))
Session_10 = Session_10.rename(columns={"Session_10": name + "_Session_10"})
```

```
    # utilisation du t l phone dans la session 11
    Session_11 = df[["Start_Date", "Session_11"]].
        groupby("Start_Date").sum("Count")
    Session_11 = Session_11.assign(Session_11=list(map(
        lambda x: 1 if int(x) > 0 else 0, Session_11.
        Session_11)))
    Session_11 = Session_11.rename(columns={"Session_11
        ": name + "_Session_11"})

    #fusion des dataframe via la date (mise en index
        par les groupby)
    merge = pd.concat(
        [Session_0,Session_1,Session_2,Session_3,
            Session_4,Session_5,Session_6,Session_7,
            Session_8,Session_9,Session_10,Session_11],
        axis=1)
    #remplissage des  lments  vides avec des 0 et
        retour du dataframe
    merge = merge.fillna(0)
    return merge

#fonction permettant de r cup rer les  vnements de
    knowledgec, puis de les agr ger
def KnowledgeC_Events(Start,End,path) :
    print("Start␣KnowledgeC␣events")
    #requ te SQLite
    print(path)
    db = sqlite3.connect(path)
    df = pd.read_sql_query('SELECT␣ZOBJECT.Z_PK,␣(
        ZOBJECT.ZSTARTDATE␣+␣978307200)␣as␣"Start_ts",␣
        datetime((ZOBJECT.ZSTARTDATE)␣+␣978307200,"
        unixepoch")␣as␣"Start_dt",␣(ZOBJECT.ZENDDATE␣+␣
        978307200)␣as␣"End_ts",datetime((ZOBJECT.
        ZENDDATE)␣+␣978307200,"unixepoch")␣as␣"End_dt",␣
        (ZOBJECT.ZENDDATE-ZOBJECT.ZSTARTDATE)␣as␣"
        Duration",␣ZOBJECT.ZSTREAMNAME␣as␣"Name",␣
        ZOBJECT.ZVALUESTRING␣as␣"Value",␣ZOBJECT.
        ZVALUEDOUBLE␣as␣"ValueDouble"␣FROM␣ZOBJECT␣ORDER
        ␣BY␣Start_ts',
                    db)
```

```
db.close()

#nettoyage du dataframe
df = set_time_and_sort(df)

#cr ation des sessions
df = set_sessions(df)

#r cup ration des  vnements  ayant eu lieu durant la p riode d'utilisation
df = df.loc[(df.Start_Date>=Start) & (df.End_Date<=End)]

#r cup ration des  vnements  "/display/orientation"
#si le dataframe est vide, r cup re le dataframe de base pour les  vnements  non ponctuels, sinon r cup re le dataframe des donn es agr g es
df_orientation = df.loc[(df.Name=="/display/orientation")]
if len(df_orientation) ==0 :
    df_orientation = get_default_df_knowledge_non_ponctual(Start, End,"orientation")
else :
    df_orientation = anonimize_knowledge_dataframe_non_ponctual_events(df_orientation,"orientation")

# r cup ration des  vnements  "/device/isPluggedIn"
# si le dataframe est vide, r cup re le dataframe de base pour les  vnements  non ponctuels, sinon r cup re le dataframe des donn es agr g es
df_plugged = df.loc[(df.Name=="/device/isPluggedIn")]
if len(df_plugged)==0 :
    df_plugged = get_default_df_knowledge_non_ponctual(Start,
```

```
        End,"isPlugged")
else :
    df_plugged =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_plugged,"isPlugged")

# r cup ration des   vnements   "/display/
   isBacklit"
# si le dataframe est vide, r cup re le dataframe
    de base pour les   vnements   non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_isbacklit = df.loc[(df.Name=="/display/isBacklit
   ")]
if len(df_isbacklit)==0 :
    df_isbacklit =
        get_default_df_knowledge_non_ponctual(Start,
        End,"isBacklit")
else :
    df_isbacklit =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_isbacklit,"isBacklit")

# r cup ration des   vnements   "/device/isLocked
   "
# si le dataframe est vide, r cup re le dataframe
    de base pour les   vnements   non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_islocked = df.loc[(df.Name=="/device/isLocked")]
if len(df_islocked) == 0 :
    df_islocked =
        get_default_df_knowledge_non_ponctual(Start,
        End,"isLocked")
else :
    df_islocked =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_islocked,"isLocked")

# r cup ration des   vnements   "/system/
   airplaneMode"
```

```
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_airplane = df.loc[(df.Name=="/system/
   airplaneMode")]
if len(df_airplane)==0:
    df_airplane =
        get_default_df_knowledge_non_ponctual(Start,
        End,"airplaneMode")
else :
    df_airplane =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_airplane,"airplaneMode")

# r cup ration des  vnements  "/wifi/connection
   "
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  ponctuels, sinon
   r cup re le dataframe des donn es agr g es
df_wifi = df.loc[(df.Name=="/wifi/connection")]
if len(df_wifi)==0 :
    df_wifi = get_default_df_knowledge_ponctual(
        Start,End,"wifi")
else :
    df_wifi =
        anonimize_knowledge_dataframe_ponctual_events
        (df_wifi,"wifi")

# r cup ration des  vnements  "/bluetooth/
   isConnected"
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_bluetooth = df.loc[(df.Name=="/bluetooth/
   isConnected")]
if len(df_bluetooth) == 0 :
    df_bluetooth =
        get_default_df_knowledge_non_ponctual(Start,
        End,"Bluetooth")
```

```
else :
    df_bluetooth =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_bluetooth ,"Bluetooth")

# r cup ration des   vnements   "/device/
   batterySaver"
# si le dataframe est vide, r cup re le dataframe
    de base pour les   vnements   ponctuels, sinon
   r cup re le dataframe des donn es agr g es
df_batterysaver = df.loc[(df.Name=="/device/
   batterySaver")]
if len(df_batterysaver)==0:
    df_batterysaver =
        get_default_df_knowledge_ponctual(Start,End,
        "batterySaver")
else :
    df_batterysaver =
        anonimize_knowledge_dataframe_ponctual_events
        (df_batterysaver,"batterySaver")

# r cup ration des   vnements   "/audio/
   outputRoute"
# si le dataframe est vide, r cup re le dataframe
    de base pour les   vnements   non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_audiooutput = df.loc[(df.Name=="/audio/
   outputRoute")]
if len(df_audiooutput)==0:
    df_audiooutput =
        get_default_df_knowledge_non_ponctual(Start,
        End,"audioOutput")
else :
    df_audiooutput =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_audiooutput,"audioOutput")

# r cup ration des   vnements   "/audio/
   inputRoute"
# si le dataframe est vide, r cup re le dataframe
```

```
    de base pour les  vnements  non ponctuels ,
   sinon r cup re le dataframe des donn es
   agr g es
df_audioinput = df.loc[(df.Name=="/audio/inputRoute
   ")]
if len(df_audioinput)==0:
    df_audioinput =
        get_default_df_knowledge_non_ponctual(Start,
        End,"audioInput")
else :
    df_audioinput =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_audioinput,"audioInput")

# r cup ration des  vnements  notification
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  notifications,
   sinon r cup re le dataframe des donn es
   agr g es
df_notificationusage = df.loc[(df.Name=="/
   notification/usage")]
if len(df_notificationusage) == 0 :
    df_notificationusage =
        get_default_df_knowledge_notification(Start,
        End)
else:
    df_notificationusage =
        anonimize_knowledge_notification(
        df_notificationusage)

# r cup ration des  vnements  "/device/
   lowPowerMode"
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  non ponctuels,
   sinon r cup re le dataframe des donn es
   agr g es
df_lowpowermode = df.loc[(df.Name=="/device/
   lowPowerMode")]
if len(df_lowpowermode)==0:
    df_lowpowermode =
        get_default_df_knowledge_non_ponctual(Start,
```

```
        End,"lowPowermode")
else :
    df_lowpowermode =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_lowpowermode,"lowPowermode")

# r cup ration des  vnements  batterie
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  batterie et siri,
    sinon r cup re le dataframe des donn es
   agr g es
df_batterypercentage = df.loc[(df.Name=="/device/
   batteryPercentage")]
if len(df_batterypercentage) == 0:
    df_batterypercentage =
        get_default_df_knowledge_percentage_siri(
        Start,End,"batteryPercentage")
else :
    df_batterypercentage =
        anonimize_percentage_and_siri(
        df_batterypercentage,"batteryPercentage")

# r cup ration des  vnements  siri
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  batterie et siri,
    sinon r cup re le dataframe des donn es
   agr g es
df_siri = df.loc[(df.Name=="/siri/ui")]
if len(df_siri)==0:
    df_siri =
        get_default_df_knowledge_percentage_siri(
        Start,End,"siri")
else :
    df_siri = anonimize_percentage_and_siri(df_siri
        ,"siri")

# r cup ration des  vnements  "/media/
   nowPlaying"
# si le dataframe est vide, r cup re le dataframe
    de base pour les  vnements  non ponctuels,
   sinon r cup re le dataframe des donn es
```

```
    agr  g  es
df_mediaplaying = df.loc[(df.Name=="/media/
   nowPlaying")]
if len(df_mediaplaying)==0 :
     df_mediaplaying =
        get_default_df_knowledge_non_ponctual(Start,
        End,"mediaPlaying")
else :
     df_mediaplaying =
        anonimize_knowledge_dataframe_non_ponctual_events
        (df_mediaplaying,"mediaPlaying")

# r cup ration des  vnements  "/app/usage"
# si le dataframe est vide, r cup re le dataframe
     de base pour les  vnements  ponctuels et pour
     les  vnements  d'utilisation d'applications,
# sinon r cup re les dataframes des donn es
    agr  g  es
df_appusage = df.loc[(df.Name=="/app/usage")]
if len(df_appusage)==0 :
     df_appusage_1 =
        get_default_df_knowledge_app_usage(Start,End
        ,"appUsage")
     df_appusage_2 =
        get_default_df_knowledge_ponctual(Start,End,
        "appUsage")
else :
     df_appusage_1 = anonimize_app_usage(df_appusage
        ,"appUsage")
     df_appusage_2 =
        anonimize_knowledge_dataframe_ponctual_events
        (df_appusage,"appUsage")

# r cup ration des  vnements  "/app/inFocus"
# si le dataframe est vide, r cup re le dataframe
     de base pour les  vnements  ponctuels et pour
     les  vnements  d'utilisation d'applications,
# sinon r cup re les dataframes des donn es
    agr  g  es
df_appinfocus = df.loc[(df.Name=="/app/inFocus")]
if len(df_appinfocus)==0 :
```

```
        df_appinfocus_1 =
           get_default_df_knowledge_app_usage(Start,End
           ,"appInfocus")
        df_appinfocus_2 =
           get_default_df_knowledge_ponctual(Start,End,
           "appInfocus")
    else :
        df_appinfocus_1 = anonimize_app_usage(
           df_appinfocus,"appInfocus")
        df_appinfocus_2 =
           anonimize_knowledge_dataframe_ponctual_events
           (df_appinfocus,"appInfocus")

    #fusion de tous les dataframe via la date (qui a
          t    mise en index par les groupby)
    merge = pd.concat(
        [df_orientation,df_plugged,df_isbacklit,
           df_islocked,df_airplane,df_wifi,df_bluetooth
           ,df_batterysaver,df_audioinput,
         df_audiooutput,df_notificationusage,
            df_lowpowermode,df_batterypercentage,
            df_siri,df_mediaplaying,df_appusage_1,
         df_appusage_2,df_appinfocus_1,df_appinfocus_2
            ],
        axis=1)

    #remplissage des     lments     nuls avec des 0
    merge = merge.fillna(0)
    #Mise en index de la date
    merge = merge.reset_index()
    merge = merge.rename(columns={"Start_Date" : "Date"
       })
    merge = merge.set_index("Date")

    #enregistrement dans un csv et retour du dataframe
    merge.to_csv("knowledge.csv")

    return merge

#fonction permettant de r cup rer le dataframe de
   base pour les donn es li es au lockdown
```

```
def get_default_df_lockdown(Start,End):
    #r cup ration du dataframe contenant les jours d'
        utilisation
    df = create_default_dataframe(Start,End)

    #cr ation de la colonne de 0
    df = df.assign(Starting_up_nb=[0 for x in df.Date])

    #mise en index de la date et retour du dataframe
    df = df.set_index("Date")
    return df

#fonction permettant de r cup r er les donn es
    li es au lockdown, de calculer les champs manquants
    et d'agr ger les donn es
def Lockdown(Start,End,path):
    #si le chemin vaut 0, la base de donn es est
        indisponible et le dataframe de base pour le
        lockdown est retourn
    if path == 0 :
        return get_default_df_lockdown(Start,End)

    day = []
    count = []
    f = open(path)
    #ouvre le fichier lockdown et parcourt les lignes
    for x in f.readlines() :
        #s'il y a "main: Starting Up" dans la ligne,
            r cup re la date et ajoute 1 au compteur
        if "main: Starting Up" in x :
            y = x.split(" ")[0] + " " + x.split(" ")
                [1].split(".")[0]
            z = datetime.strptime(y,"%m/%d/%y %H:%M:%S"
                )
            day.append(z)
            count.append(1)
        else :
            pass
    #cr    un dataframe avec toutes les informations
        sur les mises en route du t l phone
    data_lockdown = {"Date_dt": day, "Count": count, "
```

```
        Date_ts": list(map(lambda x : x.timestamp(),day)
        )}
    df = pd.DataFrame(data_lockdown)

    #si le dataframe l   est vide, retourne celui de
        base pour les donn es lockdown
    if len(df) == 0  :
        return get_default_df_lockdown(Start,End)

    #calcule la date et l'heure pour chaque   vnement
    df = df.assign(Date=list(map(lambda x: datetime.
        fromtimestamp(int(x)).date(), df.Date_ts)),
         Time=list(map(lambda x: datetime.fromtimestamp(
             int(x)).time(), df.Date_ts)))

    # si le dataframe ne contient aucun   lment
        durant la p riode d'utilisation, retourne celui
         de base pour les donn es lockdown
    if len(df.loc[(df.Date>=Start) & (df.Date<=End)])
        == 0  :
        return get_default_df_lockdown(Start,End)

    #nombre d'allumages du t l phone
    df_anonimized = df.loc[(df.Date>=Start) & (df.Date
        <=End)][["Date","Count"]].groupby("Date").sum("
        Count")
    df_anonimized = df_anonimized.rename(columns={"
        Count" : "Starting_up_nb"})

    #enregistre le dataframe et le retourne
    df_anonimized.to_csv("lockdown.csv")
    return df_anonimized

if __name__ == '__main__':
    #Date du d but de l'utilisation au format "dd.mm.
        yy     hh:mm:ss"
    Start = datetime.strptime("06.04.22␣  ␣00:00:00","%
        d.%m.%y␣  ␣%H:%M:%S").date()
    # Date de fin de l'utilisation au format "dd.mm.yy
           hh:mm:ss"
    End = datetime.strptime("06.04.22␣  ␣23:59:59", "%d
```

```
    .%m.%y   %H:%M:%S").date()
#id messenger de l'utilisateur si messenger est
    install  , sinon 0
#chemin du dossier contenant les diff rents
    dossiers dans lesquels sont stock es les
    fichiers et bases de donn es
Base_path = ""
# m me chemin qu'avant mais en brut
Brut_base_path = r""
#id de l'utilisateur
id_user = 2
#id du t l phone
id_phone = 2


#appelle toutes les fonctions qui r cup rent les
    donn es, calcule les champ manquants et
    agr gent les donn es
#il faut donner la date de d but d'utilisation, la
     date de fin d'utilisation et le chemin des
    bases de donn es
#si la base de donn es n'est pas disponible ou n'
    existe pas, il faut mettre 0
#pour certaines fonctions, il faur donner des
    param tres en plus (ud_messenger ou liste des
    mails)
print("d but du programme : " + str(datetime.now()
    ))
kn = KnowledgeC_Events(Start,End,Base_path + "/
    knowledgeC.db")
lo = Lockdown(Start,End,Base_path + "/lockdownd.log
    ")


#fusionne tous les dataframe (via la date)
merge = pd.concat([kn,lo],axis=1)

#rempli les   lments   nuls de 0
merge = merge.fillna(0)

#enregistre le dataframe final dans un fichier csv
```

```
    merge.to_csv("anonimized_data_user_" + str(id_user)
        + "_phone_" + str(id_phone) +"_2.csv")
    print("fin programme : " +str(datetime.now()))



#Auteur : Michelet Ga tan
#Adapted by Hannes Spichiger
```

Appendix E

Sc2: Code used for behavioural biometrics analysis

This Python script was used in Chapter 5 to generate an LR from the behavioural characteristics extracted using the script in Annex D. This script is partially based on (Michelet, 2021).

The script is available at `https://github.com/HSpichig/Thesis/blob/main/BB_Sc2_LR.py`

```
# import des modules d'int r t
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import math
import random
import matplotlib.pyplot as plt
import csv
from fitter import Fitter

# pip install pandas, sklearn

# groupe contenant les variables syst me
features_system = ['orientation_1_nb', '
    orientation_0_nb', 'orientation_1_duration',
                    'orientation_0_duration', '
                        orientation_1_duration_mean',
                    'orientation_0_duration_mean', '
```

```
   isPlugged_1_nb', 'isPlugged_0_nb'
   ,
'isPlugged_1_duration', '
   isPlugged_0_duration', '
   isPlugged_1_duration_mean',
'isPlugged_0_duration_mean',
'isBacklit_1_nb', 'isBacklit_0_nb',
   'isBacklit_1_duration',
'isBacklit_0_duration', '
   isBacklit_1_duration_mean',
 'isBacklit_0_duration_mean', '
    isLocked_1_nb', 'isLocked_0_nb',
'isLocked_1_duration', '
   isLocked_0_duration',
'isLocked_1_duration_mean', '
   isLocked_0_duration_mean',
'airplaneMode_1_nb', '
   airplaneMode_0_nb', '
   airplaneMode_1_duration',
'airplaneMode_0_duration', '
   airplaneMode_1_duration_mean',
'airplaneMode_0_duration_mean', '
   wifi_event_nb', '
   wifi_event_duration',
'wifi_event_duration_mean', '
   Bluetooth_1_nb', 'Bluetooth_0_nb'
   ,
'Bluetooth_1_duration', '
   Bluetooth_0_duration',
'Bluetooth_1_duration_mean', '
   Bluetooth_0_duration_mean',
'batterySaver_event_nb', '
   batterySaver_event_duration',
'batterySaver_event_duration_mean',
   'audioInput_1_nb',
'audioInput_0_nb', '
   audioInput_1_duration', '
   audioInput_0_duration',
'audioInput_1_duration_mean', '
   audioInput_0_duration_mean',
'audioOutput_1_nb', '
```

```
                        audioOutput_0_nb', '
                          audioOutput_1_duration',
                        'audioOutput_0_duration', '
                          audioOutput_1_duration_mean',
                        'audioOutput_0_duration_mean', '
                          Hidden_nb', 'Receive_nb', '
                          Dismiss_nb',
                        'Orb_nb', 'IndirectClear_nb', '
                          DefaultAction_nb', '
                          lowPowermode_1_nb', '
                          lowPowermode_0_nb',
                        'lowPowermode_1_duration',
                        'lowPowermode_0_duration', '
                          lowPowermode_1_duration_mean',
                        'lowPowermode_0_duration_mean', '
                          batteryPercentage_nb', 'siri_nb',
                        'mediaPlaying_1_nb', '
                          mediaPlaying_0_nb', '
                          mediaPlaying_1_duration',
                        'mediaPlaying_0_duration', '
                          mediaPlaying_1_duration_mean',
                        'mediaPlaying_0_duration_mean', '
                          appUsage_Session_0',
                        'appUsage_Session_1', '
                          appUsage_Session_2', '
                          appUsage_Session_3',
                        'appUsage_Session_4', '
                          appUsage_Session_5', '
                          appUsage_Session_6',
                        'appUsage_Session_7', '
                          appUsage_Session_8', '
                          appUsage_Session_9',
                        'appUsage_Session_10', '
                          appUsage_Session_11', '
                          appUsage_event_nb',
                        'appUsage_event_duration', '
                          appUsage_event_duration_mean',
                        'Starting_up_nb']


# Returns the id of element within lischte
```

```
def get_list_id(lischte, element):
    for i in range(len(lischte)):
        if lischte[i] == element:
            return i

    return -1


# Function calculating the center of gravity of a list
   of vectors
def cent_of_gravity(vector_list):
    weight = len(vector_list)  # Number of vectors to
       calculate the center of gravity of. Used for
       weighting
    dim = len(vector_list[0])  # dimension of the
       passed vectors
    result = [0.0] * dim  # instantiate a null-vector
       with dimension dim

    for v in vector_list:
        for i in range(dim):
            result[i] += v[i] / weight

    return result


# Function importing vectors from data file csv. Takes
   subfolder path from base path as input and returns
# array of vectors
def get_vectors(subpath):
    vectors = []

    with open(Base_path + subpath) as file_name:
        file_read = csv.reader(file_name)
        array = list(file_read)

    index_list = []

    for k in features_system:
        index_list.append(get_list_id(array[0], k))
```

```
    for k in array[1:]:
        turn = []
        for j in index_list:
            turn.append(k[j])
        vectors.append(turn)

    #print(vectors)
    return vectors


# Normalises over all vectors in play.
# Accepts as an input a list of a list of vectors. The
   vectors will be assembled together and normalised.
# Returns a list of a list of normalised vectors
   grouped and ordered the same way as the input
def normalise(list_o_list):
    input_structure = []
    vector_list = []
    output = []
    for i in list_o_list:
        input_structure.append(len(i))
        vector_list += i
    sc = StandardScaler()
    transformed = sc.fit_transform(X=vector_list).
       tolist()
    for i in input_structure:
        m = []
        for j in range(i):
            m.append(transformed.pop(0))
        output.append(m)
    return output


# Calculate length of a vector
def vector_len(vector):
    elements_sum = 0
    for i in vector:
        i
        elements_sum += float(i) * float(i)
    return math.sqrt(elements_sum)
```

```
# Calculates the vector distance between two vectors
def vector_distance(v1, v2):
    dim = len(v1)
    if dim == len(v2):
        distance_vector = [0.0] * dim
        for i in range(dim):
            distance_vector[i] = v1[i] - v2[i]
        return vector_len(distance_vector)
    else:
        print("Vector dimension is not equal")
        return -1


# Gets the difference from the comparison vector to the
    center of gravity of the reference
# Takes as an input a list of vectors (reference) and a
    single vector (comparison)
# Returns the distance as a float
def get_value(reference, comparison):
    cog = cent_of_gravity(reference)
    return vector_distance(cog, comparison)


def get_dist_values_intra(pop, sample_size):
    values_list = []
    for i in range(sample_size):
        vector_selection = random.choices(pop, k=15)
        random.shuffle(vector_selection)
        values_list.append(get_value(vector_selection
            [:-1], vector_selection[-1]))
    return values_list


# Conducts a PCA on the inputted values (list_o_list)
   and returns the n_c first coordinates.
def pca_transform(list_o_lists, n_c):
    list_o_lengths = []
    master_list = []
    for i in list_o_lists:
        list_o_lengths.append(len(i))
```

```
        master_list += i
    pca = PCA(n_components=n_c)
    temp = list(pca.fit_transform(master_list))
    output = []
    for i in list_o_lengths:
        m = []
        for j in range(i):
            m.append(temp.pop(0))
        output.append(m)
    return output


def get_dist_values_inter(pop1, pop2, sample_size):
    values_list = []
    for i in range(sample_size):
        vector_selection = random.choices(pop1, k=14)
        vector_ref = random.choices(pop2, k=1)
        random.shuffle(vector_selection)
        values_list.append(get_value(vector_selection,
            vector_ref[0]))
    return values_list

samp = 10000  # Number of samples for the creation of
    the reference data.
plotting = True
fitting = 0

Base_path = ""  # Path of folder containing anonymized
    Data



# Recover vectors from csv file
v_Pop1 = get_vectors("/User1_Ref/
    anonimized_data_user_1_phone_1_2.csv")
v_Pop2 = get_vectors("/User2_Ref/
    anonimized_data_user_2_phone_2_2.csv")
v_DoI = get_vectors("/Phone1_DoI/
    anonimized_data_user_1_phone_1_2.csv")  # Vector of
    values for the day of interest
v_DoI_2 = get_vectors("/Phone2_DoI/
    anonimized_data_user_1_phone_2_2.csv")
```

```
N_Pop1 = len(v_Pop1)  # Number of days in reference
    population 1
N_Pop2 = len(v_Pop2)  # Number of days in reference
    population 2

normalised = normalise([v_Pop1, v_Pop2, v_DoI, v_DoI_2
    ])
pcad = pca_transform(normalised, 1)

Pop1_intra = get_dist_values_intra(pcad[0], samp)
Pop1_inter = get_dist_values_inter(pcad[0], pcad[1],
    samp)
Pop2_intra = get_dist_values_intra(pcad[1], samp)
Pop2_inter = get_dist_values_inter(pcad[1], pcad[0],
    samp)
dist_DoI_Pop1 = get_value(pcad[0], pcad[2][0])
dist_DoI_Pop2 = get_value(pcad[1], pcad[2][0])

dist_DoI2_Pop1 = get_value(pcad[0], pcad[3][0])
dist_DoI2_Pop2 = get_value(pcad[1], pcad[3][0])

print(dist_DoI_Pop1)
print(dist_DoI_Pop2)
print(dist_DoI2_Pop1)
print(dist_DoI2_Pop2)


if fitting:
    f1 = Fitter(Pop1_intra, distributions='beta')
    f1.fit()
    print("Pop 1")
    print(f1.summary())
    print(f1.get_best())

    f2 = Fitter(Pop2_intra, distributions='expon')
    f2.fit()
    print("Pop 2")
    print(f2.summary())
    print(f2.get_best())
```

```
# titre
if plotting:
    plt.title("Distance from center Pop 1")
    plt.subplot(211)
    plt.hist(Pop1_intra, fc=(0, 0, 0, 0.2), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P1 intra")
    if fitting:
        f1.plot_pdf()
    plt.hist(Pop1_inter, fc=(0, 0, 0, 0.5), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P1 inter")
    #point du premier utilisateur (appartenant au
        groupe)
    #b = plt.axvline(new_v_y,color='g',linestyle='
        dashed', label="Groupe")
    # # point du deuxi me utilisateur (n'appartenant
        pas au groupe)
    #c = plt.axvline(new_v_z,color='r',linestyle='
        dashed', label="Autre")
    # #ajout des l gendes
    #plt.legend(handles=[b,c],bbox_to_anchor=(0.8,0.95)
        )
    #axes nomm s
    plt.plot(2 * [dist_DoI_Pop1], [0.00, 0.5], color='
        black', label="E_S1")
    plt.plot(2 * [dist_DoI2_Pop1], [0.00, 0.5], color='
        black', linestyle='dashed', label="E_S2")

    plt.legend()

    plt.ylabel("Occurrence")

    plt.subplot(212)
    plt.hist(Pop2_intra, fc=(0, 0, 0, 0.5), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P2 intra")
    plt.hist(Pop2_inter, fc=(0, 0, 0, 0.2), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P2 inter")
    if fitting:
```

```
        f2.plot_pdf()
    plt.plot(2 * [dist_DoI_Pop2], [0.00, 1.5], color='
       black', label="E_S1")
    plt.plot(2 * [dist_DoI2_Pop2], [0.00, 1.5], color='
       black', linestyle='dashed', label="E_S2")
    plt.xlabel("Distance")
    plt.ylabel("Occurrence")
    plt.legend()


    plt.show()



# Auteur : Spichiger Hannes
# Adapted from Michelet Ga tan
```

Appendix F

Sc3 & 4: Code used for GPS-Analysis

This Python script was used in Chapters 6 and 7 to generate the probabilities for the GPS evidence. The script is available at `https://github.com/HSpichig/Thesis/blob/main/BB_Sc4_LR.py`

```
import pandas as pd
from math import *
import matplotlib.pyplot as plt
from fitter import Fitter
from pyproj import Geod
from scipy.stats import t
import numpy as np




# Transforms coordinates of the form [long, lat] to
    distance to reference point (d) and angle from north
     (phi) in rad
# Takes as input the coordinates of the reference point
     (zero) and of the point to transform (coords)
# Both points are to be formatted as follows [long, lat
    ]
# Returns [d, phi] of coords
def transform_to_rad(zero, coords):
    g = Geod(ellps='WGS84') # Initiate Geode based on
        WGS84
    a, phi, d = g.inv(zero[1], zero[0], coords[1],
```

```
        coords[0])
    #print(phi)
    return [d, radians(phi)]


def read_reference_data(path):
    report_raw = pd.read_excel(path).values.tolist()
        [1:]
    report = []
    for i in range(len(report_raw)):
        id = report_raw[i][0]
        name = report_raw[i][1]
        lat = report_raw[i][9]
        long = report_raw[i][10]
        if i < len(report_raw) - 1:
            if (report_raw[i+1][9] != lat) | (
                report_raw[i+1][10] != long):
                # Remove entries where location was not
                    updated
                report.append([id, name, long, lat])
    return report


# Function returning the probability of E given P
# Inputs:
# P: Coordinates of point P in form [long, lat]
# E: Coordinates of point E in form [long, lat]
# hw: half of the wedge size used to calculate the
    angular probability
# Ref: list of reference measures from point P in form
    [[x0,y0],...]
# Returns list of probabilities in form [pphi, pdist]
    where
# pphi: the probability to observe a measurement in a
    (2 * hw) wedge centered around the angle of the
    evidence
# pdist: the probability to observe a measurement at
    the distance of the evidence given it is within the
    given wedge
def get_probabilities(P, E, Ref, hw):
    E_rad = transform_to_rad(P,E)
```

```
    Ref_rad = []

    for i in Ref:
        Ref_rad.append(transform_to_rad(P, [float(i[2])
            , float(i[3])]))

    phi_c = 0 # Initialise counter for measurement
       points within angle
    wedge = [] # Initialise list of distances within
       the wedge
    for i in Ref_rad:
        if (i[1] >= E_rad[1] - hw) & (i[1] <= E_rad[1]
           + hw):
            wedge.append(i[0])
            phi_c += 1

    print(wedge)
    if wedge:
        f = Fitter(wedge, distributions='t')
        f.fit()
        t_par = f.get_best()['t']
        pd = t.pdf(E_rad[0], t_par['df'], t_par['loc'],
             t_par['scale'])
    else:
        pd = 0
    return [float(phi_c) / float(len(Ref_rad)), pd]


# Press the green button in the gutter to run the
    script.
if __name__ == '__main__':

    hw = pi/6 # Half of the wedge size used for the
        angular probability

    P1_measurements = read_reference_data('Report_P1.
       xlsx')
    print("P1: " + str(len(P1_measurements)))
    P2_measurements = read_reference_data('Report_P2.
       xlsx')
    print("P2: " + str(len(P2_measurements)))
```

```
P1 = [6.575116326, 46.521954786] # Coordinates of
    P1
P2 = [6.573827788, 46.521598142] # Coordinates of
    P2

E = [6.5750922, 46.5219326] # Coordinates of E

e_p1 = transform_to_rad(E, P1)
e_p2 = transform_to_rad(E, P2)
print("E -> P1" + str(e_p1))
print("E -> P2" + str(e_p2))

x_val = []
y_val = []
for i in P1_measurements:
    x_val.append(float(i[2]))
    y_val.append(float(i[3]))

#plt.scatter(x_val, y_val, s=15, c='darkgray',
    marker='x', label='P1 reference')

x_val = []
y_val = []
for i in P2_measurements:
    x_val.append(float(i[2]))
    y_val.append(float(i[3]))

#plt.scatter(x_val, y_val, c='dimgray', s=15,
    marker='x', label='P2 reference')
#plt.scatter(P1[0], P1[1], c='darkgray', s=7, label
    ='P1')
#plt.annotate("P1", (P1[0]+0.00002, P1[1]), c='
    darkgray')
#plt.scatter(P2[0], P2[1], c='dimgray', s=7, label
    ='P2')
#plt.annotate("P2", (P2[0]+0.00002, P2[1]), c='
    dimgray')
#plt.scatter(E[0], E[1], c='black', s=7, label='E1
    ')
#plt.annotate("E1", (E[0]-0.0001, E[1]-0.00005), c
```

```
        ='black')

    #plt.legend()

    P1_probs = get_probabilities(P1, E, P1_measurements
        , hw)
    P2_probs = get_probabilities(P2, E, P2_measurements
        , hw)

    print("Angular probability given P1: " + str(
        P1_probs[0]))
    print("Angular probability given P2: " + str(
        P2_probs[0]))
    print("Distance probability given P1 and phi1: " +
        str(P1_probs[1]))
    print("Distance probability given P2 and phi2: " +
        str(P2_probs[1]))


    #E_P1 = transform_to_rad(P1, E)
    #plt.plot([P1[0], P1[0] + 1 * E_P1[0] * sin(E_P1
        [1]-hw)], [P1[1], P1[1] + 20 * E_P1[0] * cos(
        E_P1[1]-hw)], c='darkgray')
    #plt.plot([P1[0], P1[0] + 1 * E_P1[0] * sin(E_P1
        [1]+hw)], [P1[1], P1[1] + 20 * E_P1[0] * cos(
        E_P1[1]+hw)], c='darkgray')
    #E_P2 = transform_to_rad(P2, E)
    #plt.plot([P2[0], P2[0] + 1 * E_P2[0] * sin(E_P2
        [1]-hw)], [P2[1], P2[1] + 1.2 * E_P2[0] * cos(
        E_P2[1]-hw)], c='dimgray')
    #plt.plot([P2[0], P2[0] + 1 * E_P2[0] * sin(E_P2
        [1]+hw)], [P2[1], P2[1] + 1.2 * E_P2[0] * cos(
        E_P2[1]+hw)], c='dimgray')

    #plt.show()


#Author: Spichiger Hannes
```

Appendix G

Sc4: Code used for behavioural biometrics analysis

This Python script was used in Chapter 7 to generate the probabilities for the behavioural characteristics extracted using the script in Annex D. This script is partially based on (Michelet, 2021).

The script is available at `https://github.com/HSpichig/Thesis/blob/main/BB_Sc4_LR.py`

```
# import des modules d'int  r  t
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
import math
import random
import matplotlib.pyplot as plt
import csv
from fitter import Fitter
from sklearn.decomposition import PCA

# pip install pandas, sklearn

# groupe contenant les variables syst me
features_system = ['orientation_1_nb', '
    orientation_0_nb', 'orientation_1_duration',
                   'orientation_0_duration', '
                      orientation_1_duration_mean',
                   'orientation_0_duration_mean', '
```

```
    isPlugged_1_nb', 'isPlugged_0_nb'
    ,
 'isPlugged_1_duration', '
    isPlugged_0_duration', '
    isPlugged_1_duration_mean',
 'isPlugged_0_duration_mean',
 'isBacklit_1_nb', 'isBacklit_0_nb',
    'isBacklit_1_duration',
 'isBacklit_0_duration', '
    isBacklit_1_duration_mean',
  'isBacklit_0_duration_mean', '
     isLocked_1_nb', 'isLocked_0_nb',
 'isLocked_1_duration', '
    isLocked_0_duration',
 'isLocked_1_duration_mean', '
    isLocked_0_duration_mean',
 'airplaneMode_1_nb', '
    airplaneMode_0_nb', '
    airplaneMode_1_duration',
 'airplaneMode_0_duration', '
    airplaneMode_1_duration_mean',
 'airplaneMode_0_duration_mean', '
    wifi_event_nb', '
    wifi_event_duration',
 'wifi_event_duration_mean', '
    Bluetooth_1_nb', 'Bluetooth_0_nb'
    ,
 'Bluetooth_1_duration', '
    Bluetooth_0_duration',
 'Bluetooth_1_duration_mean', '
    Bluetooth_0_duration_mean',
 'batterySaver_event_nb', '
    batterySaver_event_duration',
 'batterySaver_event_duration_mean',
    'audioInput_1_nb',
 'audioInput_0_nb', '
    audioInput_1_duration', '
    audioInput_0_duration',
 'audioInput_1_duration_mean', '
    audioInput_0_duration_mean',
 'audioOutput_1_nb', '
```

```
                    audioOutput_0_nb', '
                    audioOutput_1_duration',
                  'audioOutput_0_duration', '
                    audioOutput_1_duration_mean',
                  'audioOutput_0_duration_mean', '
                    Hidden_nb', 'Receive_nb', '
                    Dismiss_nb',
                  'Orb_nb', 'IndirectClear_nb', '
                    DefaultAction_nb', '
                    lowPowermode_1_nb', '
                    lowPowermode_0_nb',
                  'lowPowermode_1_duration',
                  'lowPowermode_0_duration', '
                    lowPowermode_1_duration_mean',
                  'lowPowermode_0_duration_mean', '
                    batteryPercentage_nb', 'siri_nb',
                  'mediaPlaying_1_nb', '
                    mediaPlaying_0_nb', '
                    mediaPlaying_1_duration',
                  'mediaPlaying_0_duration', '
                    mediaPlaying_1_duration_mean',
                  'mediaPlaying_0_duration_mean', '
                    appUsage_Session_0',
                  'appUsage_Session_1', '
                    appUsage_Session_2', '
                    appUsage_Session_3',
                  'appUsage_Session_4', '
                    appUsage_Session_5', '
                    appUsage_Session_6',
                  'appUsage_Session_7', '
                    appUsage_Session_8', '
                    appUsage_Session_9',
                  'appUsage_Session_10', '
                    appUsage_Session_11', '
                    appUsage_event_nb',
                  'appUsage_event_duration', '
                    appUsage_event_duration_mean',
                  'Starting_up_nb']


# Returns the id of element within lischte
```

```
def get_list_id(lischte, element):
    for i in range(len(lischte)):
        if lischte[i] == element:
            return i

    return -1


# Function calculating the center of gravity of a list
   of vectors
def cent_of_gravity(vector_list):
    weight = len(vector_list)  # Number of vectors to
       calculate the center of gravity of. Used for
       weighting
    dim = len(vector_list[0])  # dimension of the
       passed vectors
    result = [0.0] * dim  # instantiate a null-vector
       with dimension dim

    for v in vector_list:
        for i in range(dim):
            result[i] += v[i] / weight

    return result


# Function importing vectors from data file csv. Takes
   subfolder path from base path as input and returns
# array of vectors
def get_vectors(subpath):
    vectors = []

    with open(Base_path + subpath) as file_name:
        file_read = csv.reader(file_name)
        array = list(file_read)

    index_list = []

    for k in features_system:
        index_list.append(get_list_id(array[0], k))
```

```
    for k in array[1:]:
        turn = []
        for j in index_list:
            turn.append(k[j])
        vectors.append(turn)

    #print(vectors)
    return vectors


# Normalises over all vectors in play.
# Accepts as an input a list of a list of vectors. The
   vectors will be assembled together and normalised.
# Returns a list of a list of normalised vectors
   grouped and ordered the same way as the input
def normalise(list_o_list):
    input_structure = []
    vector_list = []
    output = []
    for i in list_o_list:
        input_structure.append(len(i))
        vector_list += i
    sc = StandardScaler()
    transformed = sc.fit_transform(X=vector_list).
       tolist()
    for i in input_structure:
        m = []
        for j in range(i):
            m.append(transformed.pop(0))
        output.append(m)
    return output


# Calculate length of a vector
def vector_len(vector):
    elements_sum = 0
    for i in vector:
        i
        elements_sum += float(i) * float(i)
    return math.sqrt(elements_sum)
```

```python
# Calculates the vector distance between two vectors
def vector_distance(v1, v2):
    dim = len(v1)
    if dim == len(v2):
        distance_vector = [0.0] * dim
        for i in range(dim):
            distance_vector[i] = v1[i] - v2[i]
        return vector_len(distance_vector)
    else:
        print("Vector dimension is not equal")
        return -1


# Gets the difference from the comparison vector to the
    center of gravity of the reference
# Takes as an input a list of vectors (reference) and a
    single vector (comparison)
# Returns the distance as a float
def get_value(reference, comparison):
    cog = cent_of_gravity(reference)
    return vector_distance(cog, comparison)


def get_dist_values_intra(pop, sample_size):
    values_list = []
    for i in range(sample_size):
        vector_selection = random.choices(pop, k=15)
        random.shuffle(vector_selection)
        values_list.append(get_value(vector_selection
            [:-1], vector_selection[-1]))
    return values_list


# Conducts a PCA on the inputted values (list_o_list)
    and returns the n_c first coordinates.
def pca_transform(list_o_lists, n_c):
    list_o_lengths = []
    master_list = []
    for i in list_o_lists:
        list_o_lengths.append(len(i))
```

```
        master_list += i
    pca = PCA(n_components=n_c)
    temp = list(pca.fit_transform(master_list))
    output = []
    for i in list_o_lengths:
        m = []
        for j in range(i):
            m.append(temp.pop(0))
        output.append(m)
    return output


def get_dist_values_inter(pop1, pop2, sample_size):
    values_list = []
    for i in range(sample_size):
        vector_selection = random.choices(pop1, k=14)
        vector_ref = random.choices(pop2, k=1)
        random.shuffle(vector_selection)
        values_list.append(get_value(vector_selection,
            vector_ref[0]))
    return values_list


samp = 10000  # Number of samples for the creation of
    the reference data.
plotting =0
fitting = 1


Base_path = ""  # Path of folder containing anonymized
    Data


# Recover vectors from csv file
v_Pop1 = get_vectors("/User1_Ref/
    anonimized_data_user_1_phone_1_2.csv")
v_Pop2 = []
P2_adresses = ["anonimized_data_user_2_phone_2_2.csv",
    "anonimized_data_user_1_phone_4.csv",
                "anonimized_data_user_2_phone_4.csv", "
                    anonimized_data_user_3_phone_2.csv",
```

```
                "anonimized_data_user_3_phone_3.csv", "
                   anonimized_data_user_4_phone_2.csv",
                "anonimized_data_user_4_phone_3.csv", "
                   anonimized_data_user_5_phone_1_1.csv"
                   ,
                "anonimized_data_user_5_phone_1_2.csv",
                   "anonimized_data_user_6_phone_4.csv",
                "anonimized_data_user_7_phone_4.csv"]

for i in P2_adresses:
    v_Pop2 += (get_vectors("/P2/" + i))


v_DoI = get_vectors("/Phone1_DoI/
   anonimized_data_user_1_phone_1_2.csv")  # Vector of
   values for the day of interest
v_DoI_2 = get_vectors("/Phone2_DoI/
   anonimized_data_user_1_phone_2_2.csv")

N_Pop1 = len(v_Pop1)  # Number of days in reference
   population 1
N_Pop2 = len(v_Pop2)  # Number of days in reference
   population 2

normalised = normalise([v_Pop1, v_Pop2, v_DoI, v_DoI_2
   ])
pcad = pca_transform(normalised, 10)

Pop1_intra = get_dist_values_intra(pcad[0], samp)
Pop1_inter = get_dist_values_inter(pcad[0], pcad[1],
   samp)
dist_DoI_Pop1 = get_value(pcad[0], pcad[2][0])

dist_DoI2_Pop1 = get_value(pcad[0], pcad[3][0])

print(dist_DoI_Pop1)
print(dist_DoI2_Pop1)


if fitting:
    f1 = Fitter(Pop1_intra)
```

```
    f1.fit()
    print("Pop 1")
    print(f1.summary())
    print(f1.get_best())

    f2 = Fitter(Pop1_inter)
    f2.fit()
    print("Pop 2")
    print(f2.summary())
    print(f2.get_best())

# titre
if plotting:
    plt.title("Distance from center Pop 1")
    plt.hist(Pop1_intra, fc=(0, 0, 0, 0.2), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P1 intra")
    if fitting:
        f1.plot_pdf()
    plt.hist(Pop1_inter, fc=(0, 0, 0, 0.5), edgecolor='
        black', bins=np.arange(0, 15, 0.25), density=
        True, stacked=True, label="P1 inter")
    #point du premier utilisateur (appartenant au
        groupe)
    #b = plt.axvline(new_v_y,color='g',linestyle='
        dashed', label="Groupe")
    # # point du deuxi me utilisateur (n'appartenant
        pas au groupe)
    #c = plt.axvline(new_v_z,color='r',linestyle='
        dashed', label="Autre")
    # #ajout des l gendes
    #plt.legend(handles=[b,c],bbox_to_anchor=(0.8,0.95)
        )
    #axes nomm s
    plt.plot(2 * [dist_DoI_Pop1], [0.00, 0.5], color='
        black', label="E_S1")
    plt.plot(2 * [dist_DoI2_Pop1], [0.00, 0.5], color='
        black', linestyle='dashed', label="E_S2")

    plt.legend()
```

```
    plt.ylabel("Occurrence")
    plt.xlabel("Distance")

    plt.show()

# Author : Spichiger Hannes
#Adapted from Michelet Ga tan
```

Appendix H

Sc4: Code used for Password-Analysis

These Python scripts were used in Chapter 7 to obtain the inter-variability-probabilities for the password-evidence. The first script conducts an analysis of the 10 million passwords-dump (Burnett, 2015a) counting the number of appearances of each password. The second script allows to search the that way generated statistics.

The scripts are available at `https://github.com/HSpichig/Thesis/blob/main/PW_stats.py` and `https://github.com/HSpichig/Thesis/blob/main/PW_LR_getStats.py`

```
import math
import operator


def write_to_file(dictionary, file):
    sorted_dict = sorted(dictionary.items(), key=
        operator.itemgetter(1), reverse=True)
    print("Ten most frequent passwords: ")
    for k in sorted_dict[:11]:
        print(k)
    return


filename = "10-million-combos.txt"

stats = {}
count = 1
```

```
error_count = 0

print("Starting analysis")
with open(filename, 'r') as input_file:
    for line in input_file:
        try:
            pw = line.split()[1]
            if pw in stats:
                stats[pw] += 1
            else:
                stats[pw] = 1
            if (count % 100000) == 0:
                print(str(count/100000) + '% Done')
        except:
            error_count += 1
        count += 1

print(str(error_count) + ' errors detected')
print(str(len(stats.keys())) + ' distinct passwords 
   found')

write_to_file(stats, "PW_stats.txt")
print(str(count) + ' elements analyzed')

#Author: Spichiger Hannes
```

```
password = input('Enter password you would like the 
   number of occurrences of: ')

num_of_elements = 9997987
num_of_occurrences = 0

print("Searching...")

with open('PW_stats.txt', 'r') as f:
    for line in f:
        pw_stat = line.split('\t')
        if pw_stat[0] == password:
            num_of_occurrences = float(pw_stat[1])
```

```python
                print('Password found')
                break

if num_of_occurrences == 0:
    print('Password not found')

frequency = num_of_occurrences / num_of_elements

LR = 1 / (frequency * 6.5)

print("Found " + str(int(num_of_occurrences)) + " found
    in " + str(num_of_elements) + " elements")
print("Frequency: " + str(frequency))
print("LR: " + str(LR))

#Author: Spichiger Hannes
```